DOCUMENT RESUME

ED 175 237 FL 009 161

AUTHOR Van Slype, Georges
TITLE Second Evaluation of the SYSTRAN Automatic Translation System. Final Report.
INSTITUTION Commission des Communautes Europeennes (Luxembourg).
PUB DATE 78
NOTE 130p.; Some pages may be difficult to read due to poor reproducibility of original documents; appended tables have been omitted due to small print in original document

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Agriculture; *Computational Linguistics; Data Processing; Editing; English; *French; *Information Processing; Information Retrieval; *Machine Translation; Programed Texts; *Translation; Vocabulary
IDENTIFIERS *SYSTRAN System

ABSTRACT

The machine translation system SYSTRAN was assessed for translation quality and system productivity. The test was carried out on translations from English to French dealing with food science and technology. Machine translations were compared to manual translations of the same texts. SYSTRAN was found to be a useful system of information processing, but resulting translations still required much human revision. An assessment of machine translation quality put average intelligibility at 78%. It was found that, while fewer than 5% of the words in the manual translations required revision, over one-quarter of all words in the machine translations had to be corrected. The methodology and results of the evaluation are described in detail, with much data presented in tabular form. (JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

COMMISSION
OF THE
EUROPEAN COMMUNITIES

Directorate-General
Scientific and Technical Information
and Information Management

CETIL/76/78

Luxemburg

ED175237

COMMISSION OF THE EUROPEAN COMMUNITIES

Second evaluation of the SYSTRAN Automatic Translation System

Final report

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Commission of the E.C.
Loll N. Rolling

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FL 009 161

BR 18082

by Mr. Georges Van Slype

Bâtiment Jean Monnet, Plateau du Kirchberg, Boîte postale 1907, Luxembourg – Tél 43011
Télex 3423, 3446 COMEUR LU – 2752 EURDOC LU

SUMMARY

	Page
1. Management summary	0
1.1 Terms of reference of the study	0
1.2 Significance of the report	0
1.3 Aim of the evaluation	1
1.4 Versions under consideration	2
1.41 In the first evaluation	2
1.42 In the second evaluation	2
1.5 Test conditions	3
1.6 Summary of results	4
1.61 Sample as a whole	4
1.62 Typological analysis	7
1.7 Conclusions	8
1.71 Acceptability	8
1.72 Overall quality of translation	8
1.73 Typological analysis	9
1.74 Development potential	9
1.75 Organization of post-editing	10
1.8 Recommendations	11
1.81 Marketing	11
1.82 Organization	11
1.83 Improvement of the system	11
1.84 Choice of language pairs	12
2. Methodology	12
2.1 Sampling	12
2.11 Selection of texts	12
2.12 Brief typological analysis	13
2.13 Versions under consideration	18
2.2 Evaluation criteria	19
2.21 Criteria of the first evaluation	19
2.211 List of criteria	19
2.212 Critical assessment	20
2.22 Workshop on evaluation	21
2.221 List of criteria	21
2.222 Critical assessment	21
2.23 List of criteria adopted	22
2.231 General criteria	23
2.232 By type of text	23
2.233 By length of sentences	24

Page

2.24 Definition of criteria, scale of ratings and evaluation rules 24

2.24.01 Intelligibility 24
2.24.02 Fidelity 30
2.24.03 Style 31
2.24.04 Reading time 31
2.24.05 Revision and post-editing time 31
2.24.06 Number of post-editing corrections 32
2.24.07 Source of translation errors 33
2.24.08 Analysis of the main causes of error, remedies and side effects 35
2.24.09 Acceptability to the end user 37
2.24.10 Identification of sentences and counting of words 41

2.3 Reliability of evaluation 41

2.31 Procedure adopted for the first evaluation 41
2.32 Procedure adopted for the second evaluation 42

2.4 Evaluation programme 42

2.41 Production of the seven versions of the texts 42

2.411 Routing 42
2.412 Persons involved 43
2.413 Timetable of translation and post-editing 44
2.414 Time required for coding, checking and run through computer 45
2.415 Volume of the various versions 47

2.42 Evaluation work 48

2.421 Routing 48
2.422 Persons involved 48

3. Results of evaluation 54

3.1 Overall results 54

3.11 Analysis of variation in intelligibility scores given by the different evaluators 55

3.111 Results of the evaluation of control documents 55
3.112 Method of analysis of consistency among evaluators 60
3.113 Application of the method 61
3.114 Conclusions 62
3.115 Calculation of individual intelligibility correction coefficient 64
3.116 Comments 65

3.12 Summary table of the evaluation of intelligibility, fidelity and style 66

3.121 Data 66
3.122 Comments on the presentation of data 67
3.123 Comments on the results 67

	Page
3.13 Summary table of the evaluation of intelligibility	68
3.131 Data	68
3.132 Comments on the presentation of data	70
3.133 Comments on the results	70
3.14 Summary table of number of error types corrected during post-editing and revision	71
3.141 Data on post-editing	71
3.142 Data on revision	72
3.143 Comments on the presentation of post-editing data	73
3.144 Comments on the presentation of revision data	73
3.145 Comments on the results : post-editing rate	73
3.146 Comments on the results : revision rate	75
3.15 Summary table of the time required for manual translation and post-editing	76
3.151 Data	76
3.152 Comments on the presentation of data	77
3.153 Comments on the results	77
3.16 Summary table of reading time	78
3.161 Data	78
3.162 Comments on the presentation of data	79
3.163 Comments on the results	79
3.17 Summary table of the source of errors corrected during post-editing	80
3.171 Data	80
3.172 Comments on the presentation of data	81
3.173 Comments on the results	81
3.174 Adjusted data	82
3.175 Comments on the adjusted results	82
3.18 Analytical table of causes and remedies for the main errors in machine translation	84
3.181 List of abbreviations	84
3.182 Data	84
3.183 Comments	95
3.19 Acceptability	97
3.191 Analysis of the 17 questionnaires	97
3.192 Comments on the scope of the survey	101
3.193 Comments on the results	101
3.2 Analysis of results by type of document	102
3.21 Summary table of intelligibility and post-editing rate, by source	103
3.22 Summary table of intelligibility and post-editing rate, by document type	104
3.23 Summary table of intelligibility and post-editing rate, by subject	105
3.24 Summary table of intelligibility and post-editing rate, by original languagz	106

Page

3.3 Analysis by length of sentences 108

4. Annexes 110

4.1 Tables I
4.2 Tables II } Tables omitted as unreproducible.

Second evaluation of the SYSTRAN English-French automatic translation system of the Commission of the European Communities

1. Management summary

1.1 Terms of reference of the study

The Commission of the European Communities has acquired an automatic translation system called SYSTRAN.

The first version of the system, an experimental version covering the language pair English into French was delivered in September 1976. It was tested in October 1976 and an initial evaluation was made covering both qualitative and economic aspects (our summary report of 2 September 1977).

The results of this evaluation were deemed to be encouraging enough to justify an improvement contract with the supplier of the system.

The second version of the English-French SYSTRAN, resulting from this contract, was supplied in May 1978, tested in June 1978 and a new evaluation was made dealing this time solely with the quality of the translation.

This report describes the conditions under which the second evaluation was carried out and gives the results :

- in summary form, in § 1 : Management summary
- in detail, in § 2 : Methodology, and
  in § 3 : Results

1.2 Significance of the report

The impact of this evaluation of the second version of the Commission's English-French automatic translation system, i.e. a version which can no longer be regarded as experimental, could be far-reaching.

We therefore decided to give this report the status of an "original scientific memoir" as defined by ICSU (International Council of Scientific Unions) (1)

---

(1) 'A text is regarded as belonging to the category of "original scientific memoirs' when it is written in such a way that a qualified research worker, specializing in the same branch of science, is able, simply on the basis of the information given :
a) to reproduce the experiments and secure the results described with equal accuracy or within the limits of experimental error specified by the author;
b) to repeat the author's observations and judge his findings;
c) to check the accuracy of the analysis and deductions on which the author's findings are based'.

(From Unesco publication standard, adopted by ICSU)

For this reason we have incorporated all the material used for the evaluation :

- the complete text of the 20 documents in the sample, retyped in the 7 versions under consideration : original text, automatic translation, 2 manual translations and 3 post-edited versions of the automatic translation (this text is given in a separate document, Working Document No 1)

- the complete text of the computer printout of the automatic translation and the manuscript of the three post-edited versions of the 20 documents (Working Documents No 2)

- analytical tables, sentence by sentence and document by document with evaluation ratings and comments (appended to the present report).

This means that anyone interested may review our calculations and, if necessary, make measurements based on criteria other than those we have used. Our measurements may also be combined in other ways to obtain different ratios.

Many of the problems associated with evaluating the quality of machine translation are far from being solved and an approach which is complementary to, or different from ours could prove useful. We have, in some cases, made suggestions ourselves in this report.

1.3 The aim of the evaluation

Our objectives are more extensive than in the first evaluation and cover aspects arising from the fact that the system is no longer at an experimental stage.

These objectives can be seen on two levels :

- macroevaluation

  . to provide a definitive appreciation of the quality of the second version of SYSTRAN;

  . to compare the quality of the second version with that of the first;

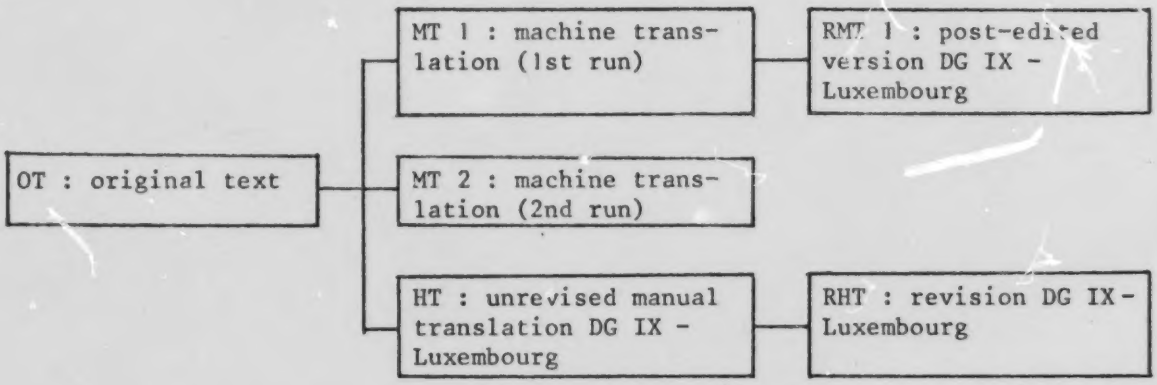
  . to make an initial assessment of the acceptability of the system to the end user;

- microevaluation

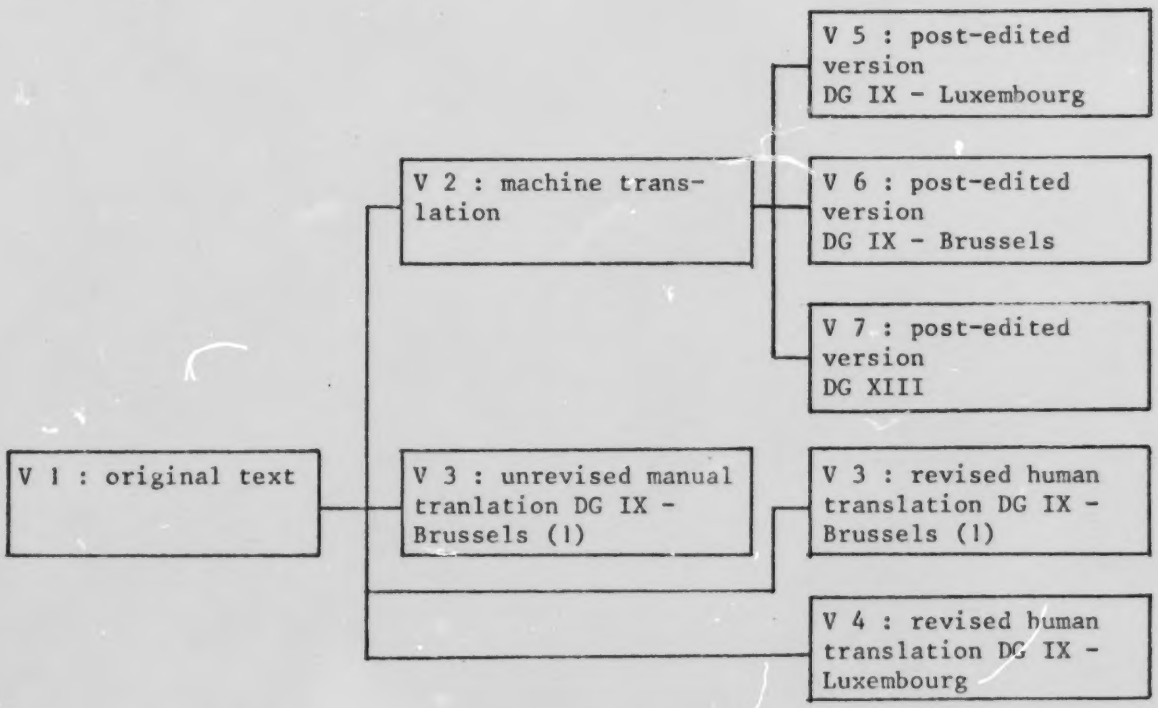
  . to analyse the possibility of improving the system at a later date.

1.4 Versions under consideration

1.41 - in the first evaluation (October 1976) (T = text)


OT : original text
MT 1 : machine translation (1st run)
RMT 1 : post-edited version DG IX - Luxembourg
MT 2 : machine translation (2nd run)
HT : unrevised manual translation DG IX - Luxembourg
RHT : revision DG IX - Luxembourg

1.42 - in the second evaluation (June 1978) (V = version)


V 1 : original text
V 2 : machine translation
V 3 : unrevised manual tranlation DG IX - Brussels (1)
V 5 : post-edited version DG IX - Luxembourg
V 6 : post-edited version DG IX - Brussels
V 7 : post-edited version DG XIII
V 3 : revised human translation DG IX - Brussels (1)
V 4 : revised human translation DG IX - Luxembourg

(1) Some of the texts supplied had been revised and others were unrevised.

1.5 Test conditions

Items	First evaluation (October 1976)	Second evaluation (June 1978)
Software	1st (experimental) version	2nd (improved) version
size	2 900 000 bytes	3 200 000 bytes
Dictionaries	25 000 words or inflexions	44 000 words or inflexions
	4 500 expressions	10 600 expressions
Number of computer runs	2 (1 after updating the dictionaries on the basis of the errors found - 700 words and 300 expressions updated)	1, without dictionary updating
Sample tested	11 200 words; grouped in 506 sentences	12 303 words, grouped in 656 sentences
subject (agriculture and food)	science and technology; administration	science and technology; administration; economics
Computer	IBM 370/145 - 3072 K, under OS/V1	IBM 370/158 - 3000 K under OS/VS1
Printing	upper case only ('poor' typography)	upper and lower case ('rich' typography)
Computer processing time :	(average 1st and 2nd runs)	(single run)
- total elapsed time	17 minutes	38 minutes
- CPU time		
. Translation	4 min. 27 sec. 96	1 min. 53 sec. 69
. preparation 'rich' typography	0	6 min. 29 sec. 13
Real translation time	2 508 words/min CPU (or 150 000 words/hour CPU)	1 468 words/min CPU or 88 000 words/h CPU)
IBM 370/145 - translation speed	2 508 words/min. CPU	489 words/min. CPU

1.6 Summary of results

1.61 Sample as a whole

Criteria	1st evaluation (October 1976)	2nd evaluation (June 1978)
- Intelligibility		
. Original text	97 %	99 %
. Machine translation without post-editing		
. . 1st run (dictionaries not updated)	45 %	78 %
. 2nd run (after updating)	75 %	-
. Revised manual translation	98 %	98 %
. Post-edited machine translation	96 %	98 %
- Accuracy of machine translation without post-editing	-	73 %
- Style of machine translation without post-editing	-	76 %
- Revision or post-editing rate		
. Manual translation	5 %	12 %
. Machine translation		
. post-edited by DG IX, Luxembourg	29 %	36 %
. post-edited by DG IX, Brussels	-	48 %
. post-edited by DG XIII	-	31 %

Criteria	1st evaluation (October 1976	2nd evaluation (June 1978)
- Translation, revision and post-editing speed (basis : number of words in original text)		
. Manual translation without revision	MT (Lux) : 182 words/hour	V3 (Brux) : 276 words/hour V4 (Lux) : 257 words/hour
. Revision of manual translation	RMT (Lux) : 665 words/hour	V4 (Lux) : 1 018 words/hour
. Post-editing of machine translation	RAT 1 (Lux) : 675 words/hour	V5 (Lux) : 339 words/hour V6 (Brux) : 226 words/hour V7 (DG XIII) : 824 words/hour
- Reading speed by end user (= evaluator) (basis : number of words in the version read)		
. Original text	-	3 658 words/hour
. Machine translation	-	3 174 words/hour
. Manual translation	-	4 979 words/hour
. Post-edited machine translation	-	4 270 words/hour

Criteria	2nd evaluation (June 1978)
- Types of errors corrected during post-editing	
. Words substituted	53 %
. Words corrected	26 %
. Words moved to a different position	9 %
. Words deleted	6 %
. Words added	6 %
TOTAL	100 %
- Causes of errors corrected during post-editing	
. Source text (ambiguity, syntactical and spelling errors)	5 %
. Acquisition	1 %
. Stylistic appreciation by the post-editor	14 %
. Translation system :	
. dictionary	50 %
. analysis	16 %
. synthesis	10 %
. miscellaneous	4 %
TOTAL	100 %
- Acceptability by end user (preliminary survey among 17 users in three institutions)	- 88 % consider that the machine translation which has not been post-edited is acceptable in certain circumstances - 71 % would be interested in receiving translation of this type : . from languages they do not know . for personal documentation and working documents . given a rapid service (half a day) . given low cost

1.62 Typological analysis

Criteria		Intelligibility (2nd evaluation - June 1978)
- Community texts		82.0 %
- non-Community texts		77.3 %
- abstracts		70.3 %
- unpublished primary texts		83.0 %
- published primary texts		80.8 %
- agricultural policy		83.7 %
- agricultural regulations		78.8 %
- agricultural economics		78.0 %
- agricultural technology		73.3 %
- food technology		81.8 %
- language of original text : likely to be English		85.3 %
- likely to be other languages (including American)		74.8 %
- language of text unknown		81.7 %
- Sentence length	Frequency	
. 1 - 5 words	13 %	86.7 %
. 6 - 10 words	10 %	83.6 %
.11 - 20 words	28 %	78.9 %
.21 - 30 words	24 %	75.8 %
.31 - 40 words	15 %	65.7 %
. over 40 words	10 %	61.8 %

1.7 Conclusions

1.71 Acceptability

The pilot test carried out in conjunction with the Translation Service of DG IX of the CEC in Brussels at the beginning of 1978 demonstrated that, in its present form, the English - French version of SYSTRAN was not acceptable to professional translators in an official translation service.

The preliminary survey carried out among a limited sample of end users from DG VI (Agriculture) of the CEC, the Centre National de la Recherche Agronomique in Versailles and the Faculty of Agronomics in Gembloux is therefore of considerable interest since it in fact shows that these is a potential market for unrevised machine translations, subject to the following conditions :

- translation from languages of which the final user has no or very little knowledge, i.e. :

  . for French-speaking scientists : all languages other than French and English

  . for Commission officials : the less widely-spoken Community languages

- translation for personal use or limited circulation of :

  . scientific and technical texts and analytical reference works, for scientists

  . working documents, for officials

- a quick (half-day) and inexpensive service.

1.72 Overall quality of translation

The SYSTRAN English-French machine translation system was substantially improved between September 1976 (provisional version, first evaluation) and May 1978 (first improved version, present evaluation), and the average intelligibility level of the translations went up from 45 % to 78 %.

It must, however, be pointed out that the post-editing rate, i.e. the relative number of corrections to be made to the translated text to make it completely intelligible and stylistically correct, has increased slightly from 29 % to 31 % or 36 % (depending on the post-editor) of the number of words translated. However, it also has to be said that the revision rate, i.e. the relative number of corrections made to a manual translation after it had been typed for the first time, went up from 5 % in the first evaluation to 12 % in the present evaluation : the texts in the sample were thus apparently more 'difficult' in the second evaluation.

Over 50 % of the post-editing rate is accounted for by the substitution of more precise words for those supplied by the computer.

This leads one to question what the priorities should be for the future development and refinement of the system. It appears that emphasis should in future be placed on improving the post-editing rate rather than the intelligibility.

The question also has to be raised as to what strategy should be adopted for the improvement of the system.

So far, contracts for the improvement of the English-French system have been based on :

- day-to-day experience of the system

- translation errors detected by qualitative methods

- subjective assessment of the quantitative importance of these errors

- subjective selection of errors for priority correction.

These contracts have covered a number of specific system components, but no real improvement strategy has been defined.

This empirical method has borne fruit since the first English-French SYSTRAN improvement contract led to considerable improvement in intelligibility, and the second contract gives rise to serious hopes of further improvements.

However, the limitations of this approach are now becoming apparent with the lack of improvement in the post-editing rate.

1.73 Typological analysis

Analysis of the quality of the translation by type of document shows that it is on the whole better in respect of :

- Community texts as composed with texts originating outside the Commission

- unabridged texts as compared with abstracts

- documents dealing with agricultural policy, food technology, agricultural regulations and economics as compared with those covering agricultural technology

- texts drawn up in English as compared with those drafted in American and those which have themselves been translated from another language.

It should nevertheless be pointed out that the differences in quality between the categories rated highest and those rated lowest are not very great and that there are not many documents in each category. Thus these differences in quality could change and the situation might even be reversed with another set of sample texts.

However, the translation quality does appear to vary much more according to the length of the sentences; machine translation could be considerably more efficient if there were no more than 25 words per sentence.

1.74 Development potential

During the first evaluation in October 1976, intelligibility increased from 45 % (before updating) to 75 % (after the dictionaries had been updated).

We concluded that two steps could be taken to improve the intelligibility of the new version of SYSTRAN :

- the dictionaries could be amplified although this would not produce the spectacular improvement achieved by updating during the first evaluation because in that case updating had been based solely on the texts in the sample being translated at the time;

- the analysis and synthesis programs could be improved.

We felt sure that intelligibility of 65 - 75 % could be achieved in the new version of SYSTRAN (p. 27, Report No 4 of 13 June 1977). This proved to be an underestimate as average intelligibility has been assessed at 78 % during the present evaluation.

On the other hand, our estimated improvement in the post-editing rate (from 29 to 20 %) proved to be wrong.

From analysis of the causes of the errors found during the present evaluation and possible remedies (Cf. § 3.18), it seems likely that the next SYSTRAN improvement contract will secure a further increase in intelligibility, bringing it, it is hoped, to 80 - 85 %. From this point on, further investment will probably become increasingly uneconomic and the decision on whether a new improvement phase should be launched will have to be made.

It will probably be difficult to improve the post-editing rate as the system is conceived at present. This should, nevertheless, be the main aim of any new improvement contract.

In view of the number of post-editing corrections involving the replacement of badly translated words, it would be worth examining, among other things, the advantages of :

- compiling specialized dictionaries in specific fields (consistent with the system logic)

- incorporating dictionaries specific to certain documents or series of documents

- increasing the number of special routines (consistent with the system logic) to deal with words which have several meanings; these routines seem to be particularly useful when a given word in the source language can be translated by various words with slightly different shades of meaning in the target language (typical case in which translation by a general word covering all or part of these shades of meaning, results in good intelligibility but a high post-editing rate).

1.75 Organization of post-editing

An analysis of the post-editing results and performances achieved by the various departments shows that the motivation and attitude of the post-editor are of prime importance :

Post-editors	Post-editing rate	Post-editing speed	Intelligibility of post-edited texts
DG IX, Brussels	48 %	226 words/h	97.8 %
DG IX, Luxembourg	36 %	339 words/h	97.1 %
DG XIII	31 %	824 words/h	98.2 %

Motivation and attitude are linked to various factors, and in particular to :

- qualification : a highly-qualified translator does not appreciate, and understandably so, making trivial and repetitive corrections.

- practice : a translator used to revising manual translations finds it difficult to adapt to the very different job of post-editor.

- understanding and acceptance of the aims of the post-editing tests, so that interest is aroused.

- workload : a generally overworked person or department is rather reluctant to take part in experimental tests unless able to devote himself fully to the task.

In sum, it would appear desirable for some consideration to be given to the job of post-editing and for suitable organizational structures to be created so as to permit post-editing work to be carried out under more operational conditions.

1.8 Recommendations

1.81 Marketing

- A preliminary survey and, if the results are positive, an in-depth study of the acceptability of the system to private translation bureaux and free-lance translators.

- An in-depth study of the acceptability of the system to its end users and of the operating arrangements.

1.82 Organization

- The creation of facilities for pilot machine translation schemes (without post-editing) to be carried out under suitable conditions (particularly as regards deadlines) on behalf of end users and, possibly, independent translators.

- The creation of a structure permitting the post-editing of machine translations to be arranged within the Translation Departments of the Commission in conditions such that the post-editors are suitably motivated.

1.83 Improvement of the system

- Formulation of a real strategy for improvement of the system, comprising :

- review, possibly in co-operation with the designer of SYSTRAN and its main users throughout the world, of the various possible approaches with an estimate of the cost in relation to improved performance

- allocation of the resources ear marked for improving the system on the basis of the action or combination of actions involving :

  . the lowest cost

  . the greatest probability of improved performance.

1.84 Choice of language pairs

- Direct the selection of language pairs towards source languages with a low or average user spread and target languages which are widely used.

2. Methodology

Any methodology of evaluation contains four elements which will be discussed in the order listed below.

- sampling of the population studied
- choice of criteria on which the evaluation will be based
- statistical reliability
- evaluation programme

2.1 Sampling

2.11 Selection of texts

Unlike the first evaluation, in which the texts were selected by DG XIII, in the second evaluation we ourselves selected the documents.

Rather than approach the Translation Services of DG IX (which might have set strict standards by selecting documents that were difficult to translate) or DG XIII (which, as purchaser of SYSTRAN, might have selected easy texts), we asked DG VI (Agriculture, that is the field covered by the dictionaries of the Commission's English-French system) to provide us with a set of some 50 documents which were representative of the activities of its various departments from these we ourselves took a sample of 9 documents, from each of which we then selected from 2 - 4 consecutive pages forming a meaningful whole.

In addition, we selected a certain number of journals from the DG VI library : two journals of scientific and technical abstracts, two primary technical journals and four primary economic journals. From these we took 11 documents consisting either of the first pages of long articles, technical abstracts or economic bulletins. The abstracts and bulletins were put in groups of two or three to make "document" units of a sufficient size to allow those involved to measure the time taken for translation, post-editing and evaluation, in a sufficiently reliable manner.

In all, the sample consists of :

- 20 documents (9 Community documents and 11 non-Community) (in administrative, technical and economic fields)
- that in 656 sentences with 12 303 words (i.e. 18.75 words per sentence).

It should be noted that the sample in the first evaluation consisted of 506 sentences and 11 200 words (22.13 words per sentence).

2.12 Brief typological analysis

A list of the 20 selected documents is given below with a brief analysis based on :

- average number of words per sentence
- source
- type of document
- field and subject covered
- source language (in the case of a document in English which is itself a translation from another language) or language of text author writing in English altough his mother tongue may be another language).

This description serves as the basis for a typology which is then used in an analysis of the results of the evaluation by document type.

The problem of the typology of texts for translation has already been the subject of much discussion :

- some people question the value of such a typology in a Translation Service where the texts for translation display extremely varied and seemingly random characteristics (frequently within a single text)
- others wonder what would be taken as the basis for such a typology :

  sentence length, full-text or summaries, scientific, technological, economic, administrative or legal fields, working documents or publications, minutes, reports, speeches, journal articles, legislation or regulations .....
  with the added difficulty that many texts may pass from one "stage" to another : a report may become an article or a regulation, ...
- finally, others raise the question of the validity of such a typology, given a translation system which is itself developing and constantly evolving.

The brief typological analysis made here does not claim to give definitive answers to these questions, but to provide a certain number of concrete data as food for further thought.

Doc. No	Title	Number of			Source	Document type	Field	Subject	Source language or author's language
		Sentences	Words	Words per sentence					
1	Market and price policy (extract from a communication by the Commission to the Council - COM 73 - 1850 final of 31/10/73 : Improvement of the Common Agricultural Market; pp. 9-11).	37	929	25	CEC	Communication to the Council	Agricultural economics	Agricultural policy	? English ?
2	Food engineering - food packaging (extracts from FSTA - Food Science and Technology Abstracts - Vol. 9 No 12, December 1977) (12E444 and 12F453).	16	291	18	FSTA	2 abstracts	Food technology	- pollution - packaging	
3	The consumer and the producer (extract from the Commission's 1977 Report : The agricultural situation of the Community ISBN 92-825-0021-7; pp. 97-100).	45	983	22	CEC	published report	agricultural economics	Agricultural market	?
4	Commodity technologies - beverages - cereals and bakery products (extracts from FSTA 9-12) (12G927, 12H1983 and 12M1336).	18	281	16	FSTA	3 abstracts	Food technology	- cooking - wine - pasta	English English Czech
5	Problems and proposals for improvements (extract from the Bulletin of the European Communities, Supplement 2/75 : Stocktaking of the Common Agricultural Policy; pp. 31-33).	36	860	24	CEC	Communications to the Parliament, published in the Bulletin of the European Communities	agricultural economics	agricultural policy	?
6	Fats, oils and margarine - fish and marine products (extracts from FSTA 9-12) (12N605 and 12R649)	19	301	16	FSTA	2 abstracts	Food technology	- fats - marine products	English ? English ?

Doc. No	Title	Number of			Source	Document type	Field	Subject	Source language or author's language
		Sentences	Words	Words per sentence					
7	Proposal for a Council Regulation on a common measure for forestry in certain dry Mediterranean zones of the Community (submitted to the Council by the Commission (extract from COM (78) 169 final of 24 April 1978; pp. 4-7).	65	951	15	CEC	Communication to the Council	agricultural economics	agricultural policy (forestry)	?
8	Agricultural engineering - vegetable crops, essential oil plants (extracts from Agritrop, Vol. 1, No 1,1977).	19	256	13	Agricrop	2 abstracts	agricultural technology	- irrigation - fertilizers	French ? French ?
9	Standard definitions additional to those in the regulations relating to the FADN (extract from : Farm accountancy data network for the EEC, results 1975; COM/77/548 final, Nov. 77; pp. 51-55).	46	858	19	CEC	report on survey	agricultural economics	administrative regulation (accounting plan)	?
10	Stimulant crops, medicinal plants - rubber plants (extracts from Agritrop, Vol. 1, 1977).	20	345	17	Agritrop	2 abstracts	agricultural technology	- Cultivation of medicinal plants - Cultivation of hevea	French French
11	Summary record of the meeting of the Standing Committee on Agricultural Research, Brussels 14 April 1978 (extracts from Doc. CPRA (CR) 14/6/78;pp. 1-6).	64	922	14	CEC	minutes of meeting	agricultural economics	management of agricultural research	?
12	Food microbiology - food hygiene and toxicology (extracts from FSTA 9-12) (12B73 and 12C505).	18	308	17	FSTA	2 abstracts	food technology	- food microbiology - toxicology	English ? American

Doc. No	Title	Number of			Source	Document type	Field	Subject	Source language or author's language
		Sentences	Words	Words per sentence					
13	Credit to agriculture in the EC Member States, a comparative analysis (extract from a serial study by the Commission : Information on agriculture, No 28 February 1977; pp. 3-9).	42	879	21	CEC	published study	agricultural economics	credit to agriculture	Dutch
14	Improvement of a poultry processing factory near Bradford, West Yorkshire (text of a telex concerning project UK/132/78) (355/77).	16	232	15	CEC	telex	agricultural economics	administrative regulation (file on financial aid)	?
15	The use of acetylene inhibition of nitrous oxide reductase in quantifying denitrification in soils (extract from Swedish Journal of Agricultural Research, Vol. 7 No 4, 1977; pp. 179-181).	38	877	23	Swedish Journal of Agricultural Research	journal article	agricultural technology	treatment of soil	Swedish
16	Food and marketing (extract from Agricultural Outlook, March 1978, p. 6).	16	385	24	Agricultural outlook	Bulletin	agricultural economics	price trends	American
17	Effect of vacuum packaging on microorganisms on cut-up chickens and in chicken products (extract from Journal of Food Science , Vol. 40, No 1, Jan. - Feb. 1975; p. 50).	57	975	17	Journal of Food Science	journal article	food technology	microbiology and packaging	American

Doc. No	Title	Number of			Sources	Document type	Field	Subject	Source language or author's language
		Sentences	Words	Words per sentence					
18	Argentina - oilseeds - cattle (extracts from :	19	319	17					
	- Agra Europe, March 10,1978, p. 10				Agra Europe	Bulletin	agricultural economics	prices (wheat)	English ?
	- Grain Market News, March 10, 1978, p. 2				Grain Market News	Bulletin	agricultural economics	prices (oilseeds)	American
	- Meat - Wool, Weekly summary and statistics, March 7, 1978, p. 194.				Meat - Wool	Bulletin	agricultural economics	prices (cattle)	American
19	Thermal effects on the length of sarcomeres in muscles held at different tensions (extract from Journal of Food Science, Vol. 40, No 1, Jan. - Feb. 1975; p. 24).	39	790	20	Journal of Food Science	journal article	food technology	physiology and meat	American
20	Sugar and confectionary-cereals and cereal processing - dairy products (extract from Industrial Food Research in the countries of the EC. Topics suitable for Community research projects).	26	561	22	CEC	draft research programme	agricultural economics	Research management (sugar, cereals, dairy products)	English
	TOTAL	656	12 303	18.75					

2.13 Versions under consideration

Machine translation cannot be judged in absolute terms but only in relation to :

- firstly, the original version which itself may contain errors of spelling, syntax and punctuation as well as barely intelligible sentences

- secondly, a manual translation, which, although it results from a series of subjective decisions (in the choice of words and sentence structure), is nonetheless the only framework of reference available for a comparative assessment of the quality of machine translation

- finally, the post-edited version of the machine translation, so as to permit an evaluation of the amount of work required to correct the imperfections of the machine translation and a comparison of this amount of work with the improvement in the quality of the translation.

Our first evaluation of the English-French SYSTRAN covered a system which was still at an experimental stage (unrevised software, dictionary restricted to 25 000 words or inflexions and 4 500 expressions); at that time we therefore took into account two versions of machine translation :

- the first based on the existing system

- the second, after up-dating of the dictionaries (covering 700 words and 300 expressions, for a sample of 11 200 words) to take into account errors in the dictionary detected on the initial run.

It is obvious that in practice machine translations will not always be preceded by an up-dating of the dictionaries.

However, the Commission's English-French SYSTRAN system can now be regarded as operational in the field of agriculture (second version of the software, dictionaries of 44 000 words or inflexions and 10 600 expressions).

For these two reasons, in the present evaluation we requested a single run through the computer without prior up-dating of the dictionaries.

However, as regards the revised manual translations and the post-edited versions of the machine translation, it was considered desirable to take into account not a single version, as in the first evaluation, but a certain number of versions, produced by the different departments : so that one manual translation should not be taken as the perfect version of the machine translation but that one should have a broader yardshide for comparison since two manual translations of the same text will also vary.

In addition, post-editing can be carried out from two different approaches :

- that normally adopted in a Translation Service, which consists in devoting the time required to arrive at a text of comparable quality to that of the revised manual translation (optimum quality).

- the approach which should be adopted by the post-editor and which is to take the minimum time required to obtain a correct and comprehensible text (optimum time).

In sum, account has been taken of the following seven versions :

- V 1 : original text in English

- V 2 : machine translation

- V 3 : revised manual translation - DG IX Translation Service in Brussels

- V 4 : revised manual translation - DG IX Translation Service in Luxembourg

- V 5 : machine translation post-edited by the DG IX Translation Service in Luxembourg

- V 6 : machine translation post-edited by the DG IX Translation Service in Brussels

- V 7 : machine translation post-edited by DG XIII in Luxembourg.

The texts of the seven versions of the 20 documents in the sample are given in "Working documents" (Working documents No 1 for fair copies of edited versions, Working documents No 2 for the computer printout of the machine translation and the post-edited scripts) which are appended to this report.

2.2 Evaluation criteria

In the following subsections we shall discuss :

- the criteria used in the first evaluation

- the conclusions reached by the Luxembourg workshop on the evaluation of automatic translation systems

- the list of criteria adopted for the present evaluation

- the definition of criteria, scales of ratings and evaluation rules.

2.21 Criteria of the first evaluation

2.211 List of criteria

In our first evaluation, for groups of criteria were adopted :

- the intelligibility of each of the versions : original text (OT), manual translation before (MT) and after (RMT) revision, machine translation before post-editing, i.e. before (AT1) and after (AT2) up-dating of the dictionaries, and machine translation after post-editing and before up-dating of the dictionaries (RAT1).

  Intelligibility was defined as a subjective evaluation of the degree of clarity and comprehensibility of each sentence, on a scale from 0 to 3 :

  0 = bad or non-existent

  1 = fairly bad

  2 = fairly good

  3 = good

- the comparative amount of revision of manual translations and post-editing of machine translations (the latter before up-dating of the dictionaries) : the ratio of the total number of changes (words substituted, moved to a different position, corrected, excluded or added) to the total number of words in the original texts.

- translation of structures : recognition of sentence structures and their appropriate transposition into French by the automatic translation system : verb phrases, noun phrases, subject phrases (subjects and complements), noun and adjective complements, verb complements, predicates.

- translation of words : recognition of the various parts of speech, transposition into French and grammatical agreement : verbs, negations, nouns, substantive expressions, articles, adjectives and adjectival expressions, pronouns, and pronominal expressions, prepositions and conjunctions, invariable forms and punctuation marks.

2.212 Critical assessment

The experiment demonstrated that only the first two criteria were of any real value :

- intelligibility because it is of direct concern to the user, the latter above all requiring a text that he can read and understand

- the amount of revision and post-editing, because they are of direct interest to the Translation Service; the lower the amount of post-editing, the closer it comes to the amount of revision and the more acceptable the work of post-editing in psychological and economic terms.

The two other criteria (recognition of sentence structure and translation of words) are of an incidental value only : they interest neither the reader nor the translator and not even those responsible for maintenance of the automatic translation system, whereas their analysis is extremely time-consuming.

It was found, however, that it would be useful to adopt three new types of criteria, which were of considerable interest :

- the fidelity of the translation, i.e. the degree to which the information contained in the original text is in fact reproduced without distortion in the translation

- the relative lengths of time required to read the manual translation and the machine translation; the closer the latter is to the former, the more acceptable the translation will be to the user. By reading time we mean the time required to read and understand the text (or to become aware that it is unintelligible)

- the source of errors, i.e. a breakdown of the errors corrected at the post-editing stage into a certain number of categories, thus enabling those responsible for maintenance of the system to assess how the quality of the machine translation can be improved and to decide where priority should be given to improvement in order to bring about the greatest increase in quality. As we have pointed out

above, the categories "sentence structure" and "parts of speech" are irrelevant to system maintenance.

2.22 Workshop on evaluation

2.221 List of criteria

On 28 February 1978, the Commission of the European Communities organized a workshop in Luxembourg on evaluation in automatic translation systems.

This workshop brought together more than 50 experts from all the countries of the Community and the United States.

It was unfortunately not possible to reach any consensus on the definition of criteria or the methodology of evaluation.

The workshop nevertheless provided an opportunity to exchange ideas and above all to collect a vast body of data (more than 30 contributions) on the methodology of automatic translation evaluation.

On 1 March 1978, a select group, composed of 11 experts, met to draw conclusions from the symposium on the basis of which to formulate recommendations for the second evaluation of the Commission's English-French SYSTRAN system.

The group examined our proposals (a summary of which is given in § 2.212) and adopted them, although insisting on

- use of the Carroll 9-point scale to measure intelligibility, rather than our own 4-point scale
- the evaluation procedure (two or more evaluators, rather than one only : cf. § 2.3)
- the value of drawing up a scale of ratings for the analysis of errors, which could be used by those responsible for system maintenance.

2.222 Critical assessment

The Carroll intelligibility scale (illustrated in § 2.24.01.2) has undoubted advantages :

- its use in several evaluations of machine translation
  - . since that culminating in the celebrated ALPAC report by the National Academy of Sciences (which concluded that machine translation was of little value)
  - . and for the SYSTRAN evaluations
- valid psychometric principles, based on equal-appearing intervals : each of the points on the scale is equidistant from both the previous and subsequent points in terms of the subjective differences observed
- it can be quickly learned and is easy to apply, giving a relatively high degree of consistency between evaluators.

An examination of this scale, however, reveals one major drawback which, to our surprise, was mentioned in none of the contributions to the Luxembourg Workshop (several of which described the scale in question), and which we have discovered ourselves when using the system : in practice this scale measures both intelligibility (which is its purpose) and style (which is not its avowed objective !); it seems to us to be part of a classical culture ("Something that has been well thought out will be well expressed ...") which is ingrained in our thought patterns, but which the majority of contemporary authors (both scientific and administrative) take very little into account.

However, a text may be perfectly intelligible, although written in a poor style (score of 7 or 6 on the Carroll scale) or even quite incorrect (score 6 or 5).

In addition, machine translation frequently results in errors of style.

In view of this, to apply the Carroll scale to machine translation seems to us scientifically unjustified since it provides a biased and downgraded picture of what it claims to measure.

It should be added that this scale from 1 to 9 and not 0 to 10, subjectively distorts the value of any score; we are used to assessments either in absolute values, or in percentages, or relative to a decimal scale.

When we see, for instance, a score of 6/9, we tend to mentally translate it by 2/3 or 67 % or 6.7/10. Whereas, on a scale of 1 to 9, a score of 6/9 is in fact worth 62.5 %. Similarly, the score of 3/9 is equal to 25 % and not 33 %.

For these reasons, we decided to use two scales to evaluate intelligibility in the second evaluation :

- the Carroll scale, from 1 to 9 points

- our own scale, from 0 to 3 points (developed by the BMvD and used in our first evaluation of SYSTRAN).

In both cases, the scores obtained have been converted into percentages. An analysis of the results (§ 3.12) shows that, as was to be expected, intelligibility on the Carroll scale is systematically lower than on our own, in the case of relatively intelligible sentences; it is, however, systematically higher than our own scale for relatively unintelligible sentences (score 1, 2 or 3 on Carroll and 0 on our own).

2.23 List of criteria adopted

In the light of the experience gained from the first evaluation, the information gathered during the Luxembourg Workshop of 28 February 1978 and the recommendations made by the Working Party of 1 March 1978, the following criteria were adopted for the present evaluation :

2.231 General criteria (i.e. for the whole of the sample)

- intelligibility according to the Carroll scale of 1 to 9 points (sentence by sentence, in all versions)

- intelligibility according to the Bureau Marcel van Dijk scale of 0 to 3 points (sentence by sentence, in version 2 only : machine translation)

- fidelity according to our scale of 0 to 3 (sentence by sentence, in version 2)

- style according to our scale of 0 to 3 (sentence by sentence, in version 2)

- reading time (document by document, for all versions)

- revision and post-editing time (document by document, for the fully revised version of the manual translation (V 4) and for the three post-edited versions of the machine translation : V 5, V 6 and V 7)

- number of post-editing corrections, with breakdown by type of error : number of words substituted, corrected, moved to a different position, excluded or added (sentence by sentence, for versions 5 and 7); or total number of corrections (sentence by sentence, for version 6); division of the number of corrections by the number of words in the machine translation to arrive at the amount of revision

- total number of corrections at the revision stage (sentence by sentence, for version 4) and amount of revision

- source of errors corrected during post-editing (sentence by sentence, for version 5, based on statistics of the number of post-editing corrections)

- analysis of the causes of the main errors detected in the sentences least well translated, and of possible remedies (sentence by sentence, based on version 2 and the evaluation of its intelligibility)

- acceptability to the end user (for version 2 of all texts).

2.232 By type of text

Intelligibility and rate of post-editing (document by document, for version 2) as a function of :

- document source (CEC and non-CEC)

- document type (abstracts, primary documents for internal use, published primary documents)

- discipline : agricultural policy, agricultural regulations, agricultural economics, agricultural technology, food technology

- language of text : English or likely to be English, other languages, author's language unknown.

2.233 By length of sentences

Intelligibility (document by document, for version 2).

2.24 Definition of criteria, scale of ratings and evaluation rules

2.24.01 Intelligibility

2.24.01.1 Definition

Subjective evaluation of the degree of clarity and comprehensibility of each sentence.

The sentence is placed in its context. The evaluator is given only one version of the text to be assessed and cannot therefore refer to the original text or to a better translation. The scoring is carried out sentence by sentence.

2.24.01.2 Scale of ratings according to Carroll

Cotes	D e f i n i t i o n s
9.	Perfectly clear and intelligible. Reads like ordinary text : has no stylistic infelicities.
8.	Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or midly unusual word usage that could, nevertheless, be easily "corrected".
7.	Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.
6.	The general idea is almost immediately intelligible, but full comprehension is distincly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Post-editing could leave this in nearly acceptable form.
5.	The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands Poor word choice, grotesque syntactical arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.
4.	Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactical arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated.
3.	Generally unintelligible; it tends to real like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.
2.	Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.
1.	Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

2.24.01.3 Scale of ratings according to Bureau Marcel van Dijk

Rating	Charateristics	Comments
3	Very intelligible	The message is fully comprehensible, even if there are stylistic and/or spelling errors, and if certain words are missing, badly translated or not translated, but with a form close to the French (e.g. intelligibility)
2	Fairly intelligible	The essential part of the message has been expressed
1	Barely intelligible	Less than 50 % of the contents of the message is inteiligible
0	Unintelligible	None or virtually none of the message is comprehensible

2.24.01.4 Evaluation rules using the Carroll scale

(Text of the instruction sheet given to the evaluators)

- Procedure to be followed

The 22 texts enclosed, numbered from 1 to 22, each consist of 1 to 5 pages (1).

These texts exist in different versions : original texts in English, manual translations in French, crude machine translations, post-edited machine translations.

For each of the 22 texts, you have been given only one version to evaluate.

You are requested to examine each sentence in turn, re-reading them, if necessary two or three times and, in each case

- to assess its intelligibility
- to give an intelligibility rating using the 9 point scale of ratings on the pink sheet enclosed
- to mark the score on the record slip opposite the number of the sentence and the initial letters of the first three words in that sentence.

Before looking at each text, please enter the following details on the corresponding record slip : you name, the date and time to the minute. After entering the score for the last sentence, fill in the exact time of completing the work before going on to the next text.

After completing your evaluation of the intelligibility of the 22nd text, please return the complete file A (2), in the same order as you received it, to the office, before opening file B (3).

---

(1) Remarks (information not disclosed to the evaluators !) :
Texts 1 to 20 in versions 1 to 7 form the sample; each evaluator dealt with only part of this sample.

Texts 21 (V 3 : manual translation) and 22 (V 2 : machine translation) are control documents : all evaluators assessed these two documents.

(2) File A contained the documents to be evaluated according to the Carroll scale : D 1 to D 20 in V 1 to V 7 and D 21 - V 3 and D 22 - V 2.

(3) File B contained the documents to be evaluated according to the Bureau Marcel van Dijk scale : D 1 to D 20 in V 2 only.

- Methodology of evaluation

The pink scale of ratings, in nine points, makes it possible to assess the intelligibility of the message without taking the fidelity of the translation into account : a translated sentence may be perfectly intelligible while expressing an idea which is quite the contrary to that found in the original text.

You are not required to take this factor into account since it is the subject of another test carried out by different evaluators. For this reason you have not been given corresponding texts from the sample (translation of the English text, originals of the translations). This is to enable you to concentrate on intelligibility alone.

In order to ensure that the results arrived at by the different evaluators are as homogeneous as possible, we give below ten sentences not included in the test which you are asked to evaluate personally using the pink scale of ratings and entering your score beside each sentence in the column "individual intelligibility".

When you have completed this initial task, the scores awarded to the 10 sentences by the different evaluators will be written up on the blackboard and, as a group, we will analyse the divergencies and try to arrive at a consensus on the assessment of the intelligibility of each of these 10 sentences ("group" intelligibility).

Subsequently, each evaluator will deal with the enclosed 22 texts individually without concerning himself with the work of the other evaluators.

- Evaluation exercise

	Intelligibility	
	Individual	Group
1° L'agriculture revêt une importance primordiale dans les régions rurales et, en raison de leur compétence agricole, les services consultatifs peuvent apporter une contribution importante à la planification rurale et à la conservation de la nature et du paysage.		
2° On propose qu'un groupe d'étude soit établi pour considérer les formes d'expertise que les services consultatifs doivent développer pour rencontrer l'objectif spécifique de conseiller des cultivateurs, et d'autres organisations concernées par des régions rurales, sur les matières de la planification et de la conservation ruraux.		

	Intelligibility	
	Individual	Group
3° On se prévoirait à ce que le groupe d'étude recommande de communautés larges initiatives dans ces régions.		
4° La nouvelle technologie dans le domaine de la communication de la masse offre une importance particulière pour le groupe plus disadvantaged de cultivateurs qui tendent à limiter le contact avec les services consultatifs.		
5° Aujourd'hui,nous nous efforcerons de donner les messages net, puisque nous les voyons, que ce type de travail ajoute à l'image totale.		
6° Comme il est uncoloured des produits alimentaires et les produits de boisson étaient intrinsically en appelants.		
7° On propose qu'une étude pilote soit établie qui regarderait au problème dans une ou deux régions sujettes de matière (peut être production laitière ou céréales).		
8° La seconde possibilité est arranger des séminaires ou workshops sur Communauté au loin base conçue pour permettre au personnel consultatif clef d'échanger des connaissances et enregistre sur les matières du souci mutuel important.		
9° Après les deux introductifs exposes par tels experts éminents lui à côté de des chutes au groupe d'industrie pour contribuer aujourd'hui au débat général, dans quatre régions clairement définies, comme place vers le bas dans le schéma.		
10° Etablissement d'un système pilote pour le conseil de groupe, y compris une comparaison du conseil de groupe avec conseil individuel conventionnel.		

2.24.01.5 Evaluation rules based on the Bureau Marcel van Dijk scale

(Instruction sheet given to the evaluators)

The first evaluation, which covered 22 texts, was carried out using a nine-point scale which in fact involved both the intelligibility of the text and its stylistic quality.

You are now asked to carry out a second evaluation of only three texts, using the simpler four point scale (B) enclosed.

As you can see, this scale does not involve the style of the sentences but only their comprehensibility. A sentence may be perfectly comprehensible although very badly constructed (e.g. I tomorrow come), or inversely be incomprehensible despite having a perfectly grammatical construction (e.g. the wind blows the telephone).

The evaluation sheet should be completed in exactly the same way as for file A, but using scale B.

2.24.02 Fidelity

2.24.02.1 Definition

Subjective evaluation of the degree to which the information contained in the sentence of the original text is in fact reproduced without distorsion in the translation.

The evaluator clearly has to have the original text, which serves as a reference, and the translation to be evaluated. The scoring is carried out sentence by sentence.

In principle, the fidelity rating must be equal to or lower than the intelligibility rating : the unintelligible part of the message cannot, by definition, be found in the translation.

Any deviation between the intelligibility rating and the fidelity rating is due to an additional distorsion of the information which may be the result of :

- loss of information (silence) (e.g. word not translated)
- interference (noise) (e.g. word added by the system)
- distorsion due to a combination of loss and interference (e.g. word badly translated).

It is extremely difficult to make a detailed analysis of the inaccuracy of a translation since each sentence carries not only one item of information or a series of items of information but a message or series of complex messages.

This may open up an avenue of fascinating research.

2.24.02.2 Scale of ratings

3 : completely or almost completely accurate

2 : fairly accurate : more than 50 % of the original information is contained in the translation

1 : inaccurate : less than 50 % of the original information is contained in the translation

0 : completely or almost completely inaccurate.

2.24.03 Style

2.24.03.1 Definition

Subjective evaluation of the correctness of sentence style.

The evaluator obviously only has one version of the text to be evaluated, the scoring being carried out sentence by sentence.

This criterion has been introduced because it is important for a translation service, whose concern is to provide intelligible and accurate texts which also have a correct style (sometimes better than in the original !).

2.24.03.2 Scale of ratings

3 : correct style

2 : fairly correct style

1 : poor style

0 : incorrect style

2.24.04 Reading time

Time required to read and understand a text (or realize its unintelligibility), but not to memorize it.

This measurement of time could be carried out sentence by sentence but would then require an outside timekeeper or a recording device connected to a clock.

In this evaluation, reading time was assessed for the whole document, so that it could be carried out by the evaluator himself without interfering with his work (Cf. penultimate paragraph of § 2.24.01.4 'Procedure to be followed').

2.24.05 Revision and post-editing time

Time taken to go through a translation, refer to all or part of the original text, carry out terminological research and correct the translation.

This time was recorded document by document.

2.24.06 Number of post-editing corrections

2.24.06.1 Definition

Ratio of the number of words corrected to the number of words in the machine translation; this assessment is carried out sentence by sentence.

2.24.06.2 Evaluation rules

(Instruction sheet given to the evaluators).

Symbols	Definition	Comments
R	Number of words substituted	- if a groupe of words is replaced and if . the number of words inserted is equal to the number of words deleted, count the number of words deleted . the number of words inserted is different from the number of words deleted, take the higher number
C	Number of words corrected	- one, two or more corrections to a single word should be taken as one correction only - do not count corrections made in order to improve the legibility of letters
D	Number of words moved to a different position	- if a group of words is moved to a different position, count the number of words in the group as it appears . after the deletion of any words . before the addition of any words - if the order of two groups of words is reversed, count the number of words in the samller group - if a word that has been corrected or substituted is moved to a different position, it should also be counted under D
E	Number of words deleted	- do not count words deleted for the purposes of substitution or correction, so as to avoid counting errors twice
A	Number of words added	- do not count words added for the purposes of substitution, correction or change of position, so as to avoid counting errors twice

2.24.07 Source of translation errors

2.24.07.1 Definition

Evaluation of the probable cause of each of the corrections made to the machine translation by the post-editor, sentence by sentence.

2.24.07.2 Scale of ratings

(Copy of form used by the evaluator).

Analysis of the source of errors corrected during post-editing (version 5)

Document No

Sentence No	Initials	Number of words post-edited	Amendments due to source texts				Errors of acquisition	Amendments due to the translation system				Personal factors (post-editor)	
			Ambiguity	Incorrect or clumsy style	Syntactical error	Spelling		Dictionary	Analysis	Synthesis	Miscellaneous	Stylistic appreciation	Post-editing error

43

2.24.08 Analysis of the main causes of error, remedies and side effects

2.24.08.1 Definition

Evaluation not only of the causes but also of the remedies, the work required to implement the latter and the possible side effects of the main errors.

This work was requested for

- the 127 sentences in the sample (out of 656) with an intelligibility lower than 50 %

- the main error or errors in each of the sentences the correction of which was likely to bring the level of intelligibility of the sentence to above 50 %.

2.24.08.2 Scale of ratings

(Copy of the form used by the evaluator).

Analysis of the causes of the main errors detected in the sentences least well translated, and possible remedies

Document No

Sentence No	Type - sources of main errors	Possible remedies	Time		Probable new translation resulting from implementation of these remedies	Side effects	Comments
			Min	Max			

2.24.09 Acceptability to the end user

2.24.09.1 Definition

Subjective evaluation of the extent to which a translation is acceptable to its end user. This evaluation is carried out on samples of documents and not document by document or sentence by sentence.

2.24.09.2 Value of the acceptability study

The ultimate aim of any translation system is to give its users an understanding of texts which are initially available in a language they do not know or of which they have a poor knowledge.

A translation system may take one of the following forms :

- a translation department which provides a written translation, usually of similar quality to that of the source text

- a helpful individual who provides a rough oral translation, generally of mediocre quality.

The user of an automatic translation system can therefore be :

- either a translation department, which could use the text produced by the computer as a rough translation to be post-edited before being passed on to the end user.

- or the end user himself, where the latter is satisfied with a translation of inferior quality.

The pilot scheme organized during the first quarter of 1978 in co-operation with the Translation Service of DG IX of the CEC in Brussels showed that the present performance of SYSTRAN is not acceptable to professional translators working in a public organization. In fact, the routine nature and high rate of post-editing constitute a major drawback which is not offset by the fact that the "machine translation and manual post-editing" procedure is less time-consuming and expensive than that of "manual translation and revision".

At the present stage of SYSTRAN's development it was therefore more appropriate to study the acceptability of the machine translation to other categories of user :

- the end users, an initial sample of whom have been covered by this study

- private translation bureaux, which we are recommending that the Commission should contact at a later stage.

2.24.09.3 Scale of ratings

(Copy of questionnaire sent to centres taking part in the survey).

Questionnaire on the acceptability of automatic translations

1. Name of the evaluator (optional)

2. Department :

3. Basic training :

4. Type of work :

5. Field of activity :

6.

Languages (1)	reading knowledge	writing ability
Danish		
Dutch		
English		
French		
German		
Italian		

(1) insert a cross in the appropriate space

7. Does your work involve the subject dealt with in at least one of the documents in the sample which you have examined (1)

often	
sometimes	
never	

8. Do you consider that the translation of these documents is acceptable, knowing that they are produced by a computer and can be obtained in a very short period, of about half a day ?

in all cases	
in certain circumstances	
never	

for myself	
for certain of my colleagues and assistants	

9. If your answer to question 8 is : 'in certain circumstances', what are these circumstances ?

10. In what cases might this quality of translation be <u>adequate</u> for your purposes ?

11. In what cases might it be useful to you, on condition that you could subsequantly request a manual translation or a version of the machine translation corrected by a human translator ?

(1) insert a cross in the appropriate space

12. Which of the defects of the enclosed machine translations seem to you the most inconvenient ?

13. Would you be interested in having access to an automatic translation system providing texts of the quality of those submitter to you ?

In what direction(s) ?

From your mother tongue to another language	
From another language to your mother tongue	

14. If your answer is yes, what benefits would you gain from such a system ?

Thank you for your co-operation in this survey.

2.24.10 Identification of sentences and counting of words

(Instructions issued for the identification of sentences and counting the number of words).

- Sentence =

  sequence of words separated by full stops (.), semicolons (;), exclamation marks (!), question marks (?) and/or colons (:); dashes (-) should only be taken as dividing of sentences when the sequence of words which follows consists of a statement containing a verb

- Titles and sub-headings should also be considered as sentences

- Numbering of sentences :

  commence with 1 at the beginning of each 'document'

- Number of words :

  Any sequence of letters and/or figures separated by blank spaces and/or punctuation marks should be counted as a word. The English 's sould be counted as a word, as should the French l' and d'

- Footnotes :

  should be left where they are, after the last sentence on the page (even if split).

2.3 Reliability of the evaluation

2.31 Procedure adopted for the first evaluation

The first evaluation of the English-French SYSTRAN covered 7 documents (one of which consisted of 20 abstracts) containing a total of 506 sentences and 11 200 words.

The evaluation of the intelligibility of the various versions of the document was carried out by one person (social psychologist).

A rapid assessment of the reliability of this evaluation was made by asking 15 people to make a separate evaluation of 10 sentences taken from the sample of 506.

The average intelligibility score awarded by these 15 people to the set of 10 sentences was lower than the average intelligibility given by the main evaluator for the same 10 sentences, the deviation compared to the main evaluator being :

- less than 7 % for the original version, the manual translation and the post-edited version

- 33 % for the machine translations.

This comparatively substantial deviation was mainly attributed to :

- practice : the main evaluator was required to asses 506 + 10 sentences whereas the 15 control evaluators were only required to asses 10 sentences each

- the fact that the 15 evaluators were asked to examine the 10 sentences out of context, this not being the case for the main evaluator.

Nevertheless, it must be admitted that procedure followed left open the question of statistical reliability.

2.32 Procedure adopted for the second evaluation

An attempt was made to achieve a reasonable balance between the cost of the evaluation and the statistical reliability of the measurements, both of which increase with the number of samples.

It was for this reason that the statistical analysis was weighted towards measuring intelligibility and particularly that of the machine translation.

The intelligibility

- of the original version (V 1), the revised manual translations (V 3 and V 4) and the post-edited machine translations (V 5, V 6 and V 7) of the 20 documents were evaluated by a team of 7

- of the machine translation (V 2) was evaluated by the same team and, in addition, by two independent evaluators.

Further, two control documents, the first a manual translation and the second a machine translation, were submitted individually to the 9 evaluators, in order to calculate individual coefficients and the statistical reliability (Cf § 3.11).

2.4 Evaluation programme

The evaluation was carried out in two stages :

- production of the seven versions of the 20 documents

- an analysis, according to the criteria listed in § 2.23 and the methodology defined in § 2.24.

2.41 Production of the seven versions of the texts

2.411 Routing

The original texts were given to the Translation Services (DG IX in Brussels and Luxembourg) for manual translation and revision.

Each document was accompanied by an 'intellectual tasks record slip', to be completed by the translator, or reviser, so as to allow us to measure translation and revision time.

The original texts were also sent to the Computer Centre for acquisition and automatic translation.

The resulting machine translations were then sent to the Translation Services (DG IX in Brussels and Luxembourg) and to DG XIII, accompanied with the same 'record slip' for post-editing and the recording of the post-editing time.

2.412 Persons involved

The work described above was carried out by the following staff :

- Translation Service in Brussels

  . manual translation (V 3) : x persons
  . revision (V 3) : 0 person
  . post-editing (V 6) : 7 persons

- Translation Service in Luxembourg

  . manual translation (V 4) : 3 persons
  . revision (V 4) : 2 persons
  . post-editing (V 5) : 2 persons

- DG XIII

  . post-editing (V 7) : 1 person.

Comments

- The number of translators from the Brussels service involved in the project is not known; this is because those involved omitted to indicate their names on the record slip. The number of post-editors is, however, known since these were identified by a code number on the record slip.

- The manual translations carried out by the Brussels service specially for the evaluation were not revised.

- The number of staff involved is much higher than for the first evaluation, for which all the post-editing work was carried out by a single person working at an abnormally fast rate.

2.413 Timetable of translation and post-editing

Doc. No	Computer Centre		Brussels Translation Service				Luxembourg Translation Service				DG XIII	
	Delivery OT V1	Receipt AT V2	Delivery OT V1	Receipt MT V3	Delivery AT V2	Receipt PE AT V 6	Delivery OT V1	Receipt MT V4	Delivery AT V2	Receipt PE AT V 5	Delivery AT V2	Receipt PE AT V 7
1	12/5	14/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	29/6	20/6	28/6
2	12/5	14/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	28/6
3	12/5	14/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	29/6	20/6	28/6
4	12/5	14/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	28/6
5	12/5	19/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	29/6	20/6	28/6
6	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	5/7
7	12/5	19/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	29/6	20/6	5/7
8	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	5/7
9	12/5	19/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	29/6	20/6	5/7
10	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	5/7
11	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	14/7
12	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	14/7
13	12/5	19/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	5/7	20/6	14/7
14	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	14/7
15	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	14/7
16	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	-
17	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	5/7	20/6	-
18	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	5/7	20/6	-
19	12/5	19/6	12/5	16/6	21/6	5/7	18/5	14/6	20/6	29/6	20/6	-
20	12/5	19/6	12/5	(16/6)	21/6	-	18/5	14/6	20/6	5/7	20/6	-

OT = original text AT = automatic translation MT = revised manual translations
PE AT = post-edited automatic translation

Comments

The following points should be noted

- the substantial time between the delivery of the original text to the Computer Centre and receipt of the machine translations (more than one month). We have the formal assurance of the person responsible for coding (Mr PIGOTT) and of the head of system maintenance (Mr FRANCOIS) that this period was not used to carry out an up-dating of the dictionaries. The impartiality of the operation would however be even more definite if the Computer Centre could carry out acquisition of the original text and automatic translation within a period in which it would be materially impossible to up-date the dictionaries;

- the failure on the part of the post-editors to supply certain documents :

  . documents 1 - 3 - 5 - 7 - 9 - 13 and 20 by the Brussels Translation Service; the reason given was that an official French translation of these documents had already been made by DG IX

  . documents 16 to 20 by DG XIII, due to an excessive workload resulting from the holiday period.

In addition, documents 1 - 3 - 5 - 7 - 9 - 13 and 20 were not translated by the Brussels translators at the time of the evaluation, but the pre-existing translation was given to us along with the manual translations of the other documents.

For this reason the date of receipt of these documents has been recorded between brackets : they were available to us for the evaluation of intelligibility, but we were unable to record the time involved in their translation and revision.

2.414 Time required for coding, checking and run through computer

- Coding of 20 documents D 1 to D 20 : 10 h 15 min i.e. 5 minutes per 100 words.

- Checking of 20 documents D 1 to D 20 : 9 h 11 min i.e. 4 1/2 minutes per 100 words.

- Computer run (IBM 370/158 of 3000 K octets, under OS/VS 1 release 60) of 20 documents D 1 to D 20.

  Processing by the computer comprises three main phases, the second of which is preparation of printout in 'rich' typography; in the first evaluation of the English-French SYSTRAN, the printout was in 'poor' typography; it is therefore necessary to isolate the processing time specific to 'rich' typography in order to compare the two evaluations :

  . Phase 1 : translation proper : formatting of text for translation, preparation of sorts, sorts for consultation of dictionaries, analysis and synthesis : 1 minute 53 seconds 69 hundredths (CPU time, for the batch of 20 documents D 1 to D 20, containing 12 303 words)

. Phase 2 : preparation of printout in 'rich' typography : compaction of translated text (for example : "3 E" gives "é") and transfer from small to capital letters (at the beginning of sentences) : 6 minutes 29 seconds 13 hundredths (CPU time for the batch)

. Phase 3 : printout of the results on printer : 2 seconds and 90 hundredths (CPU time for the batch).

The elapsed time amounted to 38 minutes and 1 second for the first two phases and to 25 minutes for Phase 3.

The relatively long CPU time required for Phase 2 (preparation of printout in 'rich' typography) can be explained by the fact that this stage of the processing was programmed independently of the main translation program, priority being given to the speed of programming rather than to its efficiency.

It appears that an efficient program would have resulted in a CPU time of the order of 6 to 12 seconds for execution of the second phase.

CPU time for translation in the 1st evaluation in 1976 on the IBM 370/145 and in poor typography was 4 minutes 27 seconds 96 hundredths for 11 200 words, i.e. 2 508 words/CPU min.

Given that the 370/158 computer used in the present evaluation has a processing speed of on average 3 times that of the 370/145, the translation speed achieved in the 2nd evaluation must be divided by 3 if it is to be compared with that in the 1st evaluation.

This gives the following figures :

Translation speed in minute/CPU

. 1st evaluation, on 370/145 : 2 508 words/min.
. 2nd evaluation, corrected to 370/145 : 2 164 words/min.

This means a 14 % fall in translation speed, which can be explained by :

. the appreciably larger volume of the dictionary and slightly greater volume of software

. the factor of 3 allowed for the difference in speed between the 370/145 and the 270/158, which is an average and cannot be applied uniformly to all types of processing.

2.415 Volume of the various versions

In the table below : S = number of sentences
W = number of words

Doc. No	Version 1		Version 2		Version 3		Version 4		Version 5		Version 6		Version 7	
	S	W	S	W	S	W	S	W	S	W	S	W	S	W
1	37	929	37	994	36	955	36	1017	34	981	36	955	37	998
2	16	291	16	378	16	408	16	390	16	372	16	397	16	376
3	45	983	45	1111	45	1184	45	1211	47	1108	45	1184	45	1099
4	18	281	18	362	18	384	18	393	18	364	18	361	18	365
5	36	860	36	948	36	943	32	908	35	902	36	943	36	928
6	19	301	19	408	19	426	19	388	19	405	19	397	19	407
7	65	951	65	988	70	989	65	965	69	926	70	989	65	980
8	19	256	19	317	19	335	16	320	19	306	21	338	19	318
9	46	858	46	934	47	965	46	862	46	906	47	965	47	918
10	20	345	20	406	19	378	20	373	20	365	20	393	20	378
11	64	922	64	976	64	1052	62	1015	64	980	64	997	63	970
12	18	308	18	397	21	375	17	378	17	356	18	389	18	385
13	42	879	42	1043	45	1053	42	1027	42	991	45	1053	43	1047
14	16	232	16	277	16	292	16	278	16	263	16	296	16	272
15	38	877	38	1014	38	1043	38	1042	38	1010	38	1088	38	1037
16	16	385	16	509	16	535	16	526	16	515	16	526	-	-
17	57	975	57	1178	58	1195	56	1200	58	1160	58	1208	-	-
18	19	319	19	410	19	443	19	435	18	394	19	456	-	-
19	39	790	39	923	41	991	39	983	39	944	39	990	-	-
20	26	561	26	622	26	712	26	620	26	612	26	712	-	-
TOTAL	555	12,303	656	14.195	669	14,658	644	14.331	657	13,860	667	14,637	(500)	(10,478)
WORDS/ SEN-TENCES		18.75		21.64		21.91		22.25		21.10		21.94		21

2.42 Evaluation work

2.421 Routing

Versions V 1 to V 7 were retyped by our office before being submitted to the evaluators, so as to ensure that all were presented in the same way and without crossings out and alterations.

This was to avoid any bias (in intelligibility and reading time) due to :

- the bad presentation of the computer printouts

- the corrections made by revisers and post-editors

2.422 Persons involved

2.422.1 Intelligibility

The following persons were involved in the evaluation of intelligility :

No	Name	Qualifications	Position
1	Mr FRETIN	Engineer (Inst. Industr. Nord) and Master of Mathematical Sciences (University of Michigan)	Consulting Engineer with the Bureau Marcel van Dijk (BMvD)
2	Mrs GUAZZO	BA in Philology (Univ. of Brussels) and Master's degree in Linguistics (University of Montreal)	Analyst on French-English SYSTRAN for CEC
3	Mr TRAN	Doctorate in Mathematics (University of Paris VI)	Analyst Documentalist with BMvD
4	Mr LE	Bachelor of Agronomic Sciences and Master of Mathematical Sciences (University of British Columbia)	Analyst Documentalist with BMvD
5	Mr BOIS	Diploma in Chemistry (Strasbourg)	Expert on on-line documentary systems with BMvD
6	Mr RAZA	Doctorate in Mathematics (University of Paris)	Analyst Documentalist with BMvD
7	Mr DOTAN	Civil engineer (Centrale, Paris)	Analyst Documentalist with BMvD
8	Mr FEUILLAT	Civil engineer (University of Liège)	Consulting Engineer with BMvD
9	Mr Van SLYPE	Commercial Engineer, Degree in Economics (University of Brussels	Manager at BMvD

Comments

- Mrs GUAZZO was asked to take part in the evaluation despite the fact that she might be biased due to her work on the SYSTRAN programme because

  . it was considered useful to have one member of the system team taking part in the evaluation of intelligibility for psychological reasons

  . Mrs GUAZZO was the only competent person available to carry out the evaluation of the source of errors corrected at the post-editing stage and to analyse the errors in the least intelligible sentences.

  . any bias in Mrs GUAZZO's evaluation could be measured.

- Four of the nine evaluators were analysts documentalists, i.e. persons used to rapidly extracting the essential information from documents and abstracts, which might result in a certain bias which could also be measured.

The work was distributed as follows :

Evaluators 1 to 7

- Intelligibility on the Carroll scale

  . all versions (V 1 to V 7) of all documents (D 1 to D 20) of the sample; each document/version was subject to one evaluation by one evaluator only : the distribution of documents/versions between the evaluators is shown in the table on the folliwing page

Table showing the distribution of documents (D 1 to D 20) and versions (V 1 to V 7) among the seven evaluators

Score out of 9	Evaluators No						
	E 1	E 2	E 3	E 4	E 5	E 6	E 7
D 1	V 1	V 2	V 3	V 4	V 5	V 6	V 7
D 2	V 1	V 2	V 3	V 4	V 5	V 6	V 7
D 3	V 1	V 2	V 3	V 4	V 5	V 6	V 7
D 4	V 2	V 3	V 4	V 5	V 6	V 7	V 1
D 5	V 2	V 3	V 4	V 5	V 6	V 7	V 1
D 6	V 2	V 3	V 4	V 5	V 6	V 7	V 1
D 7	V 3	V 4	V 5	V 6	V 7	V 1	V 2
D 8	V 3	V 4	V 5	V 6	V 7	V 1	V 2
D 9	V 3	V 4	V 5	V 6	V 7	V 1	V 2
D 10	V 4	V 5	V 6	V 7	V 1	V 2	V 3
D 11	V 4	V 5	V 6	V 7	V 1	V 2	V 3
D 12	V 4	V 5	V 6	V 7	V 1	V 2	V 3
D 13	V 5	V 6	V 7	V 1	V 2	V 3	V 4
D 14	V 5	V 6	V 7	V 1	V 2	V 3	V 4
D 15	V 5	V 6	V 7	V 1	V 2	V 3	V 4
D 16	V 6	-	V 1	V 2	V 3	V 4	V 5
D 17	V 6	-	V 1	V 2	V 3	V 4	V 5
D 18	V 6	-	V 1	V 2	V 3	V 4	V 5
D 19	-	V 1	V 2	V 3	V 4	V 5	V 6
D 20	-	V 1	V 2	V 3	V 4	V 5	V 6
D 21	V 3	V 3	V 3	V 3	V 3	V 3	V 3
D 22	V 2	V 2	V 2	V 2	V 2	V 2	V 2
Score out of 3							
V 2	D 1	D 4	D 7	D 10	D 13	D 16	D 19
V 2	D 2	D 5	D 8	D 11	D 14	D 17	D 20
V 2	D 3	D 6	D 9	D 12	D 15	D 18	-

. in the case of the control documents D 21/V 3 (manual translation) and D 22/V 2 (machine translation) : all seven evaluators assessed both documents.

- Intelligibility on the Bureau Marcel van Dijk scale

. version 2 (machine translation)of all documents (D 1 to D 20) of the sample; the distribution between the evaluators was as shown on the table, that is one evaluation only per document.

Evaluator 8

- On the Carroll scale

. control documents D 21/V 3 and D 22/V 2

- On the Bureau Marcel van Dijk scale

. Version 2 (automatic translation) of all documents (D 1 to D 20) in the sample.

Evaluator 9

- On the Carroll scale

. Version 2 of all documents (D 1 to D 20) in the sample

. control documents D 21/V 3 and D 22/V 2

- On the Bureau Marcel van Dijk scale

. Version 2 of all documents (D 1 to D 20) in the sample.

Summary

- the machine translation of all the documents in the sample was evaluated five times (Carroll method : evaluators 1 to 7 and evaluator 9; Bureau Marcel van Dijk method : evaluators 1 to 7 and evaluators 8 and 9)

- the original text, the two revised manual translations and the three post-edited machine translations of all the documents in the sample were evaluated once (Carrol method : evaluators 1 to 7). It was in fact considered pointless to evaluate texts with a degree of intelligibility approaching 100 % several times. The fact that the evaluation was spread over seven evaluators is a sufficient guarantee of its validity.

- the two control texts were evaluated nine times (Carroll method : evaluators 1, 2, 3, 4, 5, 6, 7, 8 and 9).

2.422.2 Fidelity and style

The fidelity and style of the machine translation (V 2) of all documents D 1 to D 20 were assessed by one evaluator (E 8).

The results of this evaluation show that fidelity and style achieve

scores very similar to those of intelligibility; for this reason it was not considered necessary to have these two criteria assessed by more than one evaluator.

2.422.3 Reading, revision and post-editing time

These times were measured by the participants themselves (respectively : evaluators from our team, DG IX revisers in Luxembourg, post-editors from DG IX in Brussels and Luxembourg and from DG XIII).

2.422.4 Number of revision and post-editing corrections

Counting of the corrections was carried out as follows :

- by evaluator E 8 for

  . all documents (D 1 to D 20) of version 4 (revision by DG IX in Luxembourg), without any analysis of the type of correction.

  . all documents (D 1 to D 20) of version 5 (post-edited by DG IX in Luxembourg), with an analysis of the corrections by type (except in the case of document D 16 which contained too many corrections for the errors to be identified separately and for which only the total number of corrections was recorded)

  . all the documents available (D 2 - 4 - 6 - 8 - 10 - 11 - 12 - 14 - 15 - 16 - 17 - 18 and 19) in version 6 (post-edited by DG IX in Brussels), without analysis of the corrections by type.

- by a tenth evaluator (Mr DELVAUX, commercial engineer, Director of Studies at the Bureau Marcel van Dijk) for

  . all the documents available (D 1 to D 15) in version 7 (post-edited by DG XIII).

2.422.5 Source of errors corrected at the post-editing stage

This type of evaluation obviously requires a thorough knowledge of the SYSTRAN translation system. Consequently, with the Commission's agreement, the task was carried out by Mrs GUAZZO (analyst on the French-English SYSTRAN system), who worked on the basis of :

- our own analytical list of the number of errors by type corrected during post-editing (version 5)

- original texts, the machine translation and the manuscript of the post-edited version

- a complete edition of the English-French SYSTRAN lexicons.

Note : document D 16 was not analysed for the reason given above (§ 2.422.4).

2.422.6 Analysis of the main causes of error and appropriate remedies

Here too, a thorough knowledge of the system was required to make the evaluation.

With the Commission's agreement, we requested Mr PIGOTT (English-French system and English -French dictionary maintenance) and Mrs GUAZZO (French-English system and maintenance of French-English grammar rules) to examine the list of sentences with an intelligibility lower than 50 % which we (evaluator E 9, scale of 3) had drawn up and which contained 127 sentences of the 656 in the sample.

Initially, Mr PIGOTT and Mrs GUAZZO carried out their evaluations independently. Unfortunately, they did not have the hexadecimal output on sentence analysis but worked on the basis of either the fair copy of the machine translation (Mrs GUAZZO), or the computer printout (Mr PIGOTT).

Subsequently, the two evaluations were compared by Mr PIGOTT, who made the following observations :

"A rapid comparison of our two reports would indicate the following :

1. the errors detected are in most cases the same;

2. we have both had a tendency to pick up out the most serious errors or causes of error and have often assumed that less serious errors were not the cause of lack of intelligibility;

3. the time estimates on programming changes vary by a factor of 10 or more, a possible explanation being that Mrs GUAZZO has estimated the time necessary to modify the programs to correct the specific error under consideration whereas I felt it more reasonable to estimate the time in terms of the number of man-hours spent both on collecting sufficient context of a similar type to ensure the modification would cover as many cases as possible and on the actual program writing, up-dating and checking;

4. the time estimates on coding are comparable although in my estimate I again included related aspects such as punching, updating and correction. I do not believe either of us have attempted to estimate the time required to "spot" the errors (choice of texts, punching, post-editing, feedback, etc.)"

Finally, in the third stage of the operation, a table was drawn up incorporating the results of both evaluations. This work was done by Mrs GUAZZO and the resulting table (Cf § 3.18) was approved by Mr PIGOTT. For reasons connected with the heavy workload during the holiday period, it was only possible to carry out this work for Documents 1 to 17, which is nevertheless fairly representative of the sample.

2.422.7 Acceptability

The best persons to evaluate the acceptability of a machine translation are obviously its potential users. We therefore contacted :

- Mr ONIDI, Assistant to the Director-General for Agriculture of the Commission of the European Communities in Brussels

- Miss CAGNAC, of the CNRA (National Centre for Agronomic Research) at Versailles

- Mrs WILLAM, of the Faculté des Sciences Agronomiques de l'Etat (State Faculty of Agronomics) at Gembloux

All three kindly agreed to take part in our survey and to distribute our questionnaires (§ 2.24.09.2) to a sample of six officials (DG Agriculture) and twelve research scientists (6 at Versailles and 6 at Gembloux).

Together with the questionnaire, the participants received automatic translations which had not been post-edited and the corresponding original texts :

- of a Community nature in the case of DG Agriculture (documents Nos D 1, 3, 7, 9, 11, 13, 14, 18 and 20)

- of a scientific and technical nature in the case of Versailles and Gembloux (documents Nos D 2, 4, 6, 8, 10, 11, 12, 14, 15, 17, 19 and 20).

The following numbers of questionnaires were returned duly completed :

- 4 from DG VI of the CEC

- 10 from the CNRA

- 3 from the Faculté de Gembloux.

3. Results of the evaluation

The results will be discussed in the following order :

- the overall results for the sample as a whole

- analysis of the results by text type

- analysis of the results by length of sentence.

3.1 Overall results

Analytical results, sentence by sentence, are attached :

Table I : Intelligibility scores for each of the seven versions of the twenty documents in the sample and the number of revision corrections made to version 4 and post-editing corrections made to version 5 of the 20 documents.

Table II : Scores for intelligibility, fidelity and style and number of errors corrected by type and source in the case of version 2 : 'machine translation' of the twenty documents in the sample.

Summary figures for each document corresponding to the totals of tables I and II are presented below with comments.

3.11 Analysis of the variation in the intelligibility scores given by the different evaluators

3.111 Results of the evaluation of the control documents

Two documents (D 21 and D 22), outside the sample, were given scores by all the evaluators, in order to determine the extent to which any variation could be attributed to differences between the evaluators themselves.

The two documents in question were a manual translation (D 21 - V 3) and a machine translation which had not been post-edited (D 22 - V 2).

The scores awarded to each of the sentences in these two documents by each of the nine evaluators are shown in the following two tables.

3.111.1 Manual translation

Analysis of variation in intelligibility scores

DOCUMENT No 21 VERSION : 3 TYPE : manual translation
SCORES : 1 to 9

Sentence No	Evaluation group assessment							Individual evaluations	
	Evaluators No							Evaluators No	
	1	2	3	4	5	6	7	8	9
1	9	9	9	9	9	9	9	9	9
2	9	9	9	9	9	9	9	9	9
3	9	9	6	9	9	9	9	9	9
4	9	9	9	9	9	9	9	9	9
5	8	8	9	8	9	9	8	8	8
6	9	9	8	9	9	9	9	9	9
7	9	8	9	9	9	9	9	9	9
8	9	9	9	9	9	9	9	9	9
9	9	9	8	9	9	9	9	7	8
10	9	9	9	9	9	9	9	7	9
11	9	9	9	9	9	9	9	9	9
12	9	9	9	9	9	9	9	9	9
13	9	8	9	9	9	9	9	9	9
14	9	8	8	9	9	9	9	7	9
15	9	9	9	9	9	9	9	9	9
16	9	9	9	9	9	9	9	9	9
17	9	9	9	9	9	9	9	9	9
18	9	8	9	9	9	9	9	7	9
19	9	6	9	9	9	9	9	6	8
20	9	9	9	9	9	9	9	8	9
21	9	9	9	9	9	9	9	9	9
22	9	9	9	9	9	9	9	8	9
23	9	8	9	9	9	9	9	9	9
24	9	7	9	9	9	9	9	8	9
25	9	7	9	9	9	9	9	9	9
26	9	9	9	9	9	9	9	9	9
27	9	9	9	9	9	9	9	9	9
28	9	9	9	9	9	9	9	9	9
29	9	9	9	9	9	9	9	8	9
30	9	9	9	9	9	9	9	8	9
31	9	9	9	8	9	9	8	6	9
32	9	9	9	9	9	9	9	7	8
33	9	9	9	9	9	9	9	9	9
34	9	8	9	9	9	9	9	9	9
35	9	9	9	9	9	9	9	9	9
36	9	8	9	9	9	8	9	8	8
37	9	9	9	9	9	8	9	9	9
Total	332	318	327	331	333	331	331	310	328

Sentence No	1	2	3	4	5	6	7	8	9
Individual averages	8.97	8.59	8.84	8.95	9.00	8.95	8.95	8.38	8.86
%	100	95	98	99	100	99	99	92	98
Square deviation	0.97	18.9	11.0	1.89	0	1.89	1.89	30.7	4.32
Group average	8.89							8.62	
%	99							95	
General average	8.83								
%	98								

3.111.2 Machine translation

Analysis of variation in intelligibility scores

DOCUMENT No 22 VERSION : 2 TYPE : automatic translation
SCORES : 1 to 9

Sentence No	Evaluation group assessment							Individual evaluations	
	Evaluators No							Evaluators No	
	1	2	3	4	5	6	7	8	9
1	8	9	9	9	8	8	8	6	9
2	8	7	9	9	8	8	9	8	8
3	7	5	6	8	6	5	7	6	6
4	7	4	7	9	3	5	6	4	6
5	7	4	8	9	6	7	9	6	5
6	7	9	9	8	8	7	9	8	6
7	7	7	8	9	7	8	8	7	6
8	7	8	9	9	7	7	9	7	8
9	7	4	7	8	6	5	8	6	7
10	7	5	7	8	6	7	8	7	7
11	7	6	7	9	6	7	8	6	7
12	6	9	8	9	9	8	9	9	9
13	9	7	8	9	8	5	8	7	8
14	7	6	8	9	6	7	8	7	7
15	6	9	8	9	8	7	9	8	8
16	7	6	9	9	8	3	8	7	5
17	8	8	9	9	8	7	9	8	8
18	5	6	5	7	4	7	5	5	5
19	7	7	7	9	6	4	8	4	7
20	8	9	9	9	8	9	9	9	9
21	6	8	8	9	5	2	9	4	8
22	7	6	8	9	6	5	8	5	8
23	6	1	6	5	2	4	7	2	3
24	6	4	7	5	5	4	7	5	7
25	5	2	7	6	3	2	8	3	5
26	9	6	9	8	8	2	9	6	8
27	6	5	6	3	8	2	9	6	8
28	7	6	7	3	7	2	8	3	6
29	8	7	8	8	8	8	8	7	8
30	8	7	8	9	8	8	8	7	8
31	9	9	8	9	9	7	9	8	9
32	9	5	9	9	8	7	9	6	8
33	8	6	8	8	7	5	8	6	7
34	9	8	9	9	8	9	9	8	9
35	9	9	9	9	9	9	9	9	9
36	9	7	8	9	7	6	8	7	7
37	8	7	8	9	7	6	8	7	7
38	9	6	8	9	8	5	9	6	8
39	7	7	7	9	7	7	7	6	6
40	8	6	8	8	6	2	3	5	6

Sentence No	1	2	3	4	5	6	7	8	9
41	7	4	7	8	4	7	7	5	6
42	8	7	8	9	8	9	8	7	8
43	9	7	9	9	7	8	9	7	8
44	8	6	7	6	6	6	8	6	6
45	8	6	8	8	7	6	7	7	7
46	7	3	7	6	5	5	5	5	7
47	7	5	7	9	6	8	8	7	7
48	8	6	9	8	8	8	9	7	8
49	8	6	9	8	8	8	8	7	7
50	8	5	7	5	4	8	8	6	8
51	8	4	8	8	7	8	9	7	7
52	8	5	8	9	7	6	8	7	7
53	7	5	7	6	4	8	8	6	7
54	8	7	9	9	9	8	8	7	8
55	9	9	9	9	9	9	9	9	9
56	7	2	6	3	2	3	5	1	1
57	6	4	6	5	7	5	7	6	6
58	9	9	9	9	8	8	9	8	9
59	8	7	8	8	7	8	8	7	8
60	8	5	8	6	6	7	8	6	7
61	8	3	7	4	5	7	7	5	6
62	8	5	8	6	5	7	8	6	7
63	7	5	7	2	5	7	8	5	5
64	7	3	7	2	6	2	8	3	4
65	7	5	8	8	6	4	8	4	7
66	7	5	8	7	6	3	8	6	6
Total	495	395	514	504	434	406	530	408	462
Individual averages	7.50	5.98	7.79	7.64	6.58	6.15	8.03	6.18	7.00
%	81	62	85	83	70	64	88	65	75
Square deviation									
Group average	7.10							6.59	
%	76							70	
General average	6.98								
%	75								

3.112 Method of analysis of consistency among evaluators

The method used to analyse the variance was based on the assumption that the various evaluators (columns of tables) are similar.

The variations within any given column should, therefore, be comparable to the variations between columns, since they arise from the same random causes, independent of the evaluators.

Taking the table of n lines and m columns, the calculation method is as follows :

Source of variance	Sum of squares	Degrees of freedom	Mean square
Between columns	$S_1 = n \sum_{j=1}^{m} (\bar{x}_j - \bar{x})^2$	m-1	$\sigma_1^2 = \frac{S_1}{m-1}$
Within columns	$S_2 = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$	N-m	$\sigma_2^2 = \frac{S_2}{N-m}$

where : N = total number of cases (n x m)

$\bar{x}_j$ = average of column j

$\bar{x}$ = general average

One then caculates

$$F = \frac{\sigma 1^2}{\sigma 2^2}$$

The value arrived at for F is compared to the limit value found in any table, for the degrees of freedom indicated.

If the deviation calculated is higher than the limit found in the table, then the two variances compared are not equal and thus the assumption that the evaluators are consistent as a group does not hold good.

3.113 Application of the method

3.113.1 Manual translation

- Analysis of variance (evaluators 1 to 7) for Document 21, Version 3, Manual translation.

	Sum of squares	Degrees of freedom	Mean square
Between columns	Se = 4.50	6	$F_1 = 0.75$
Within columns	Sp = 36.54	252	$F_2 = 0.15$
Total			$F = 5$

At 5 % the limite value of F is 2.1; consequently the differences between evaluators are significant.

- Analysis of variance (evaluators 1 to 9) for document 21, Version 3, Manual translation

	Sum of squares	Degrees of freedom	Mean square
Between columns	13.05	8	$F_1 = 1.63$
Within columns	71.56	296	$F_2 = 0.24$
Total			$F = 6.75$

Conclusion : heterogeneous sample.

It should be noted that the variance and the strictness of scores awarded by evaluators 2 and 8 are different from the others.

- Analysis of variance (excluding evaluators 2 and 8) for Document 21, Version 3, Manual translation.

	Sum of squares	Degrees of freedom	Mean square
Between columns	0.74	6	F_1 = 0.12
Within columns	21.96	252	F_2 = 0.08
Total			F = ± 1.5

In this case, the remaining evaluators can be considered as behaving in the same way.

3.113.2 Machine translation

- Analysis of variance (evaluators 1 to 9) for Document 22, Version 2, Machine translation

	Sum of square	Degrees of freedom	Mean square
Between columns	327	8	F_1 = 40.87
Within columns	1 458	528	F_2 = 2.70
Total			F = 15

The evaluators therefore react differently.

3.114 Conclusions

It can be seen from these calculations of variance that the assumption that the evaluators react in the same manner is only borne out in one case : the evaluation of the manual translation (D 21 - V 3) by eliminating evaluators 2 and 8, who differ significantly from the others.

Another point worth noting as regards the behaviour of evaluators is that the variation in scores increases as intelligibility falls.

This behaviour should perhaps be evened out by decreasing the accuracy of low scores and increasing that of high scores.

Our analysis of variance also shows that while significant differences exist between evaluators, there are nonetheless outweighed by differences arising from the texts themselves.

In fact, one can compare :

- the differences between the average scores awarded by the evaluators to a given text ("average" line in the preceding tables). For the automatic translation (D 22 - V 2) this standard deviation is of the order of 0.79 of a point.

- the differences between the average scores awarded by all evaluators to all documents (Table 3.13), which, for the same version is 17 %, or 1.5 points.

We therefore attempted to homogenize, at least to some extent, the scores given by evaluators 1 to 7 (responsible for evaluating the intelligibility of the various versions of documents D 1 to D 20 of the sample) by calculating a correction coefficient for each of them. Bearing in mind the comment above on the decrease in variance with the increase in intelligibility, the correction coefficient was calculated for manual and machine translations, but applied only in the case of machine translations since the intelligibility of both manual translations and the post-edited versions was virtually 100 %.

3.115 Calculation of individual intelligibility correction coefficient

Evaluators	Scores		Individual correction coefficient		
	Manual translation (D 21 / V 3)	Machine translation (D 22 / V 2)	Manual translation	Machine translation	Verification Col.3 x Col.5
(1)	(2)	(3)	(4)	(5)	(6)
E 1 Mr FRETIN	100 %	81 %	0,99	+ 0,94	++ 76,14 %
E 2 Mrs GUAZZO	95 %	62 %	1,04	1,23	76,26 %
E 3 Mr TRAN	98 %	85 %	1,01	0,89	75,65 %
E 4 Mr LE	99 %	83 %	1	0,92	76,36 %
E 5 Mr BOIS	100 %	70 %	0,99	1,09	76,30 %
E 6 Mr RAZA	99 %	64 %	1	1,19	76,16 %
E 7 Mr DOTAN	99 %	88 %	1	0,86	75,68 %
GROUP TOTAL	690 %	533 %	-	-	532,55 %
GROUP AVERAGE	99 %	76 %	1	1	-
E 8 Mr FEUILLAT	92 %	65 %			
E 9 Mr Van SLYPE	98 %	75 %			
GENERAL TOTAL	880 %	673 %			
GENERAL AVERAGE	98 %	75 %			

+ $\frac{76}{81} = 0,94$ ++ 0,94 x 81 = 76,14 %

The averages of the scores awarded to the 37 sentences of document 21 / V 3 (Col. 2) and to the 66 sentences of document 22 / V 2 (Col. 3)were calculated for each of the participants (E 1 to E 7) in the evaluation.

The two averages (D 21 / V 3 and D 22 / V 2) were calculated for the group. The correction coefficient for each evaluator (Cols. 4 and 5) is constituted by the ratio of the group average to the individual average.

The correction which had to be made to the score for the manual translation was extremely slight and was disregarded in the following table.

However, the adjustment which had to be made to the score for the machine translations was relatively significant and was in fact applied in the tables given.

Application of this factor to document 22 / V 2, in the preceding table (Col. 6) shows that this did not affect the general average of the group.

The scores given by evaluators E 8 and E 9 for the same documents 21 / V 3 and 22 / V 2 have been provided for information only but no correction coefficient was calculated since these two evaluators assessed all documents V 2.

Evaluators 1 to 7, however, each assessed some of the versions of all documents and their correction coefficient must therefore be applied to all scores to permit the scores for each document in version 2 to be compared with the same document in the other versions.

3.116 Comments

- The score awarded by Mrs GUAZZO to the machine translation is the lowest of the 9 evaluators : this bias, which was to be anticipated, results from an over-concern to be fair ... which is all to her credit ! but which has not distorted the assessment due to application of the correction coefficient.

- The 4 analysts documentalists (E 3 - 4 - 6 and 7) gave the automatic translation an average score of 80 %; hence a bias of 4 % compared to the group average (76 %).

3.12 Summary table of the evaluation of intelligibility, fidelity and style

· Version No 2 : Machine translation

3.121 Data

DOC No	No of sentences V 2	No of words V 2	% INTELLIGIBILITY SCALE OF 3 EVALUATORS 1 - 7 GROSS SCORE	EVALUATOR	COEF.	CORRECTED SCORE	E 8	E 9	AV.	% INTELLIGIBILITY SCALE OF 9 EVALUATORS 1 - 7 GROSS SCORE	EVALUATOR	COEF.	CORRECTED SCORE	E 9	AV.	% FIDELITY E 8	% STYLE E 8
1	37	994	95	E 1	0,94	89	96	85	87	69	E 2	1,23	85	74	79,5	86	88
2	16	378	100	E 1	0,94	94	94	88	92	78	E 2	1,23	96	87	91,5	83	92
3	45	1111	94	E 1	0,94	88	87	90	88	67	E 2	1,23	82	82	82	77	79
4	18	362	80	E 2	1,23	98	78	81	86	83	E 1	0,94	78	79	78,5	76	78
5	36	948	61	E 2	1,23	75	77	81	78	80	E 1	0,94	75	69	72	75	74
6	19	408	72	E 2	1,23	89	74	79	81	81	E 1	0,94	76	70	73	74	75
7	65	988	91	E 3	0,89	81	88	88	86	92	E 7	0,86	79	78	78,5	86	78
8	19	317	88	E 3	0,89	78	79	63	73	87	E 7	0,86	75	63	69	68	70
9	46	934	86	E 3	0,89	77	70	74	74	93	E 7	0,86	80	64	72	70	75
10	20	406	78	E 4	0,92	72	73	53	66	88	E 6	1,19	100	51	75,5	70	72
11	64	976	80	E 4	0,92	74	83	81	79	69	E 6	1,19	82	68	75	77	77
12	18	397	78	E 4	0,92	72	74	72	73	63	E 6	1,19	75	69	72	74	72
13	42	1043	93	E 5	1,09	100	75	76	84	81	E 5	1,09	88	64	76	73	80
14	16	277	88	E 5	1,09	96	67	77	80	60	E 5	1,09	65	64	64,5	71	58
15	38	1014	94	E 5	1,09	100	78	66	81	70	E 5	1,09	76	60	68	76	74
16	16	509	35	E 6	1,19	42	21	25	29	22	E 4	0,92	20	25	22,5	17	29
17	57	1178	68	E 6	1,19	81	75	83	80	71	E 4	0,92	65	68	66,5	71	85
18	19	411	44	E 6	1,19	52	56	79	62	48	E 4	0,92	44	63	53,5	47	65
19	39	923	87	E 7	0,86	75	79	84	79	84	E 3	0,89	75	74	74,5	75	78
20	26	622	95	E 7	0,86	82	74	90	82	80	E 3	0,89	71	77	74	76	81
TOTAL	656	14.195	1.607			1.615	1.488	1.515	1.540	1 466			1.487	1.349	1.418	1.422	1.480
SIMPLE AVERAGE	33	710	80,4 %		1,0049	80,8 %	74,4 %	75,8 %	77 %	73,3 %		± 1	74,4	67,5 %	71 %	71 %	74 %
AVERAGE WEIGHTED BY NUMBER OF SENTENCES %			82,3		0,9915	81,6	77,1	78,5	78,8	75,7			76,2	69,0	72,6	73,8	76,3
AVERAGE WEIGHTED BY NUMBER OF WORDS %			82		0,9988	81,9	76,3	77,9	78,3	74,6			75,7	68,4	72,1	73,1	76,1

3.122 Comments on the presentation of data

- Application of the individual correction coefficient to the scores of evaluators 1 to 7 should not affect the overall total for the 20 documents. In fact, due to rounding-off, the total has been modified slightly, up or down for the simple or weighted averages respectively.

- The 3 scores underlined were purposely limited to 100. In the case of these 3 documents, application of the correction coefficient would have resulted in a score of over 100 which is meaningless !

3.123 Comments on the results

- The averages weighted by the number of sentences are slightly higher than the simple averages. This is due to the system's poorer performance in the translation of abstracts and bulletins, that is shorter documents

- The averages weighted by the number of words are not significantly different from those weighted by the number of sentences

- The % scores resulting from application of the Bureau Marcel van Dijk are significantly higher than those awarded on the Carroll scale.

  In the guidelines and the remainder of the report, account has been taken of the score awarded on our own three-point scale, because :

  . our initial evaluation (October 1976) is based on this scale and we were anxious to be able to compare the results of the two evaluations

  . the Carroll scale seems to us less appropriate than our own, for the reasons given in § 2.222.

- The scores for fidelity and style are very close to those for intelligibility.

3.13 Summary table of the evaluation of intelligibility (9-point scale)

Version 1 : original text (evaluators 1 - 7, unrevised)
Version 2 : machine translation (evaluators 1 - 7, revised)
Version 3 : manual translation DG IX, Brussels (evaluators 1 - 7, unrevised)
Version 4 : manual translation, DG IX, Luxembourg (evaluators 1 - 7, unrevised)
Version 5 : Post-edited version, DG IX, Luxembourg (evaluators 1 - 7, unrevised)
Version 6 : post-edited version, DG IX, Brussels (evaluators 1 - 7, unrevised)
Version 7 : post-edited version, DG XIII

3.131 Data

Doc. No	Version 1		Version 2		Version 3		Version 4		Version 5		Version 6		Version 7	
	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.
1	37	100	37	85	36	97	36	94	34	98	36	95	37	96
2	16	100	16	96	16	99	16	98	16	99	16	95	16	98
3	45	100	45	82	45	99	45	94	47	100	45	97	45	99
4	18	98	18	78	18	99	18	99	18	96	18	99	18	99
5	36	98	36	75	36	93	32	100	35	94	36	100	36	99
6	19	97	19	76	19	98	19	98	19	92	19	100	19	100
7	65	98	65	79	70	100	65	99	69	99	70	98	65	100
8	19	98	19	75	19	100	16	98	19	100	21	96	19	100
9	46	99	46	80	47	100	46	93	46	100	47	98	47	100
10	20	100	20	100	19	97	20	100	20	84	20	100	20	96
11	64	99	64	82	64	100	62	100	64	95	64	100	63	96
12	18	100	18	75	21	97	17	100	17	86	18	100	18	95
13	42	99	42	88	45	97	42	100	42	100	45	98	43	99
14	16	99	16	65	16	99	16	100	16	100	16	91	16	98
15	38	100	38	76	38	98	38	100	38	100	38	89	38	97

Doc. No	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	% Intel.	Number of sentences	%
16	16	100	16	20	16	99	16	98	16	95	16	100	-	-
17	57	100	57	65	58	99	56	99	58	99	58	100	-	-
18	19	98	19	44	19	100	19	98	18	100	19	100	-	-
19	39	99	39	75	41	92	39	99	39	98	39	99	-	-
20	26	97	26	71	26	100	26	100	26	92	26	99	-	-
TOTAL	656	1979	656	1487	669	1963	644	1967	657	1927	667	1954	500	1472
SIMPLE AVERAGE	33	98.9	33	74.4	33	98.1	32	98.3	33	96.3	33	97.7	33	98.1
STANDARD DEVIATION		1.05		17.29		2.23		2.18		4.74		3.13		1.73
CONFIDENCE INTERVAL 95 %		± 0.47		± 7.77		± 1.00		± 0.98		± 2.13		± 1.41		± 0.90
RANGE OF SIMPLE AVERAGES AT 95 % CONFIDENCE	98.43 to 99.37		66.63 to 82.17		97.1 to 99.1		97.5 to 99.1		94.17 to 98.43		96.29 to 99.11		97.2 to 99.0	
AVERAGE WEIGHTED BY NUMBER OF SENTENCES	99.0		76.2		98.2		98.2		97.1		97.8		98.2	
FIRST EVALUATION	97.3		45.0		98.3				96					

3.132 Comments on the presentation of data

- Documents D 1 - 3 - 5 - 7 - 9 - 13 and 20 of versions 3 and 6 are texts for which an official translation in French existed prior to the evaluation and which DG IX, Brussels, neither retranslated or post-edited; the texts given to the evaluators were therefore the same for the two versions; the scores awarded by the two evaluators to each of the identical documents are consistent to within 1 or 2 %, except in the case of D 5 (deviation of 7 %)

- In addition, for version 2 we compared the gross scores with the corrected scores; the difference is not significant :

	Gross	Corrected
Average	73.3	74.4
Standard deviation	16.6	17.3
Confidence interval 95 %	± 7.5	± 7.8

3.133 Comments on the results

- As in the first evaluation, the intelligibility of all the manual versions is close to 100 %. The deviations between weighted averages of the various versions (minimum : 97.1 %; maximum : 99 %) are marginally smaller than in the first evaluation.

  This is most probably due to the fact that 7 evaluators were involved in the case of the figures given in the previous table, as against only 1 evaluator during the first evaluation.

- The intelligibility of the 3 post-edited versions of the second evaluation is higher than that of the post-edited version in the first evaluation. In our view, this is not due to better post-editing; the explanation probably lies in the fact that the assessment of intelligibility was carried out :

  . on the manuscript of the post-edited version (computer printout with erasures) in the first evaluation

  . on a retyped fair copy in the second evaluation.

- The deviations in the intelligibility of versions V 3 to V 7 are not significant.

3.14 Summary table of number of error types corrected during post-editing and revision

3.141 Data on post-editing

Version 5 : post-edited by DG IX, Luxembourg
Version 6 : post-edited by DG IX, Brussels
Version 7 : post-edited by DG XIII

DOC NO	NUMBER OF WORDS V 2	VERSION 5							VERSION 6		VERSION 7						
		SUBST	CORR	DIFF. POS	EXCLUDED	ADDED	TOTAL	RATE %	TOTAL	RATE %	SUBST	CORR	DIFF. POS	EXCLUDED	ADDED	TOTAL	RATE %
1	994	177	76	28	43	31	355	36	-	-	143	47	17	31	29	267	27
2	378	82	46	9	15	8	160	42	191	51	55	34	4	7	8	108	29
3	1111	184	115	39	36	17	391	35	-	-	169	61	18	19	21	288	26
4	362	62	38	14	10	8	132	36	141	39	68	23	19	8	7	125	35
5	948	193	53	26	20	14	306	32	-	-	163	62	21	23	21	290	31
6	408	38	25	19	16	8	106	26	176	43	65	23	17	9	12	126	31
7	988	214	76	9	21	11	331	34	-	-	196	48	11	29	42	326	33
8	317	97	11	5	6	2	121	38	186	59	54	29	5	12	9	109	34
9	934	224	71	30	16	5	346	37	-	-	188	76	11	35	20	330	35
10	406	78	31	3	32	3	147	36	219	54	98	18	7	13	5	141	35
11	976	213	96	34	25	9	377	39	456	47	176	76	8	28	27	315	32
12	397	29	46	32	22	6	135	34	192	48	35	17	20	23	12	107	27
13	1043	247	101	39	20	16	423	41	-	-	224	48	17	28	34	351	34
14	277	83	35	10	7	7	142	51	182	66	80	16	5	9	14	124	45
15	1014	122	96	31	15	18	282	28	486	48	126	49	18	33	42	268	26
16	509	-	-	-	-	-	-		>80	-	-	-	-	-	-	-	-
17	1178	224	129	34	8	16	411	35	430	37	-	-	-	-	-	-	-
18	410	142	53	17	10	8	230	56	289	70	-	-	-	-	-	-	-
19	923	117	75	21	9	16	238	26	438	47	-	-	-	-	-	-	-
20	622	112	65	16	3	3	199	32	-	-	-	-	-	-	-	-	-
TOTAL V 5/6	7046	1287	681	229	175	109	2481	35	3386	48	-	-	-	-	-	-	-
%	100	18	10	3	2	2	35		43		-	-	-	-	-	-	-
TOTAL V 5/7	10553	2043	916	328	304	163	3754	36	-	-	1840	627	198	307	303	3275	31
%	100	19	9	3	3	2	36	-	-	-	17	6	2	3	3	31	-
TOTAL V 6/7	4535	-	-	-	-	-	-	-	2229	49	757	285	103	142	136	1423	31
%	100	-	-	-	-	-	-	-	49	-	17	6	2	3	3	31	-
TOTAL V 5	13686	2638	1238	416	334	206	4832	36.5	-	-	-	-	-	-	-	-	-
%	100	19	9	3	2	2	35	-	-	-	-	-	-	-	-	-	-

3.142 Data on revision

Version 4 : revised by DG IX, Luxembourg

Doc. No	Number of words V 4	Total number of words changed	Amount of revision
1	1,017	93	9
2	390	48	12
3	1,211	171	14
4	393	33	8
5	908	165	18
6	388	26	7
7	965	176	18
8	320	25	8
9	862	79	9
10	373	33	9
11	1,015	74	7
12	378	47	12
13	1,027	225	22
14	278	16	6
15	1,042	76	7
16	526	51	10
17	1,200	148	12
18	435	38	9
19	983	121	12
20	620	80	13
Total	14,331	1,725	12
%	100	12	-

3.143 Comments on the presentation of post-editing data

- Document No 16, the intelligibility of which was extremely low, gave rise to substantial degree of post-editing, and consequently numerous corrections. It was not possible to record the various types of correction and for this reason the document has not been included in the total;

- For versions 6 and 7, seven and five documents repectively were not post-edited. In order to arrive at comparable data we therefore calculated a series of part totals :

  . TOTAL V 5/6 : totals for documents existing in versions 5 and 6 (i.e. D 2 - 4 - 6 - 8 - 10 - 11 - 12 - 14 - 15 - 17 - 18 and 19); the totals in column V 6 are those for all documents available in V 6

  . TOTAL V 5/7 : totals for all documents existing in versions 5 and 7 (i.e. D 1 to D 15); the totals given in column V 7 correspond to all documents avaialble in V 7

  . TOTAL V 6/7 : totals for all documents available in versions 6 and 7 (i.e. D 2 - 4 - 6 - 8 - 10 - 11 - 12 - 14 and 15)

  . TOTAL V 5 : totals for all documents available in V 5 (D 1 to D 15 and D 17 to D 20).

3.144 Comments on presentation of revision data

It was only possible to record the errors revised for version 4 documents, since the majority of version 3 texts were submitted to us either unrevised, or in final typed or printed form so that it was impossible to detect the corrections.

3.145 Comments on the results : post-editing rate

- The first point worth noting is the particularly high number of post-editing corrections for version 6 : 48 %, compared to the high, but nevertheless more reasonable number for version 5 : 36 %. Is there perhaps such a thing as over-enthusiasm on the part of the post-editor ?

- Despite the considerable discrepancy between the three post-editing results (31, 36 and 48 %), the intelligibility is virtually identical (97.1, 97.8 and 98.2 %).

- The post-editing rate in all three cases is higher than that recorded during the first evaluation (29 %), despite the substantial improvement in intelligibility.

Four possible explanations can be suggested :

  . the post-editing in the first evaluation was not carried out under normal conditions : it was noted at the time that post-editing had been carried out abnormally quickly, requiring approximately the same time as the revision of the manual translation whereas the number of corrections for the latter was only 5 %.

The post-editor of DG IX, Luxembourg, in the first evaluation probably adopted the same approach as the post-editor of DG XIII during the second evaluation, with comparable results : post-editing rate 29 % and 31 %. It is nonetheless true that in its present version the system requires from 30 to 36 % post-editing.

. it would appear from the increase in the revision rate (cf. § 3.146) that the texts for translation were more difficult; nevertheless, despite this handicap, the system was able to produce a much more intelligible translation than in the first translation. Consequently, the added difficulty of the sample does not entirely explain the increase in the post-editing rate.

. even though the intelligibility of the translations has improved substantially, the system has not broken through a quality threshold which will allow a reduction in the post-editing rate. In other words : the system produces as many errors as before so that the post-editing rate is unchanged; however, each of these errors is less serious and consequently intelligibility has improved.

If this assumption were to prove correct, it is difficult to envisage how SYSTRAN, as conceived at present, could become a useful instrument for the translation of documents requiring a correct style.

. the present policy of the system's lexicographers appears to be to produce a lexicon which is as universal as possible, by giving each word in the source language the widest possible meaning in the target language. It is clear that the consequence of such a policy is to maximize the "generality" of the translation (the system will translate a specific term by a generic term with a low probability of error, rather than by a specific term with a high error probability). This approach results in generally improved intelligibility, but to the detriment of precision; if the translation is required to reproduce the precision of the original text, a large degree of post-editing is required.

This point seems to us of great importance and to merit an examination by those responsible for the system since it may lead them to redefine their objective :

.. do they require an all-round system which produces translations which are for the most part intelligible but in not very good style (rough translation), which can be used without post-editing ?

.. or do they want a system with more specific subject-bound lexicons, which can produce a translation which is for the most part intelligible and has a sufficiently correct style requiring only a low degree of post-editing (10 to 15 %) by a Translation Service, before being passed on to the user ?

These two objectives may not be incompatible as regards the translation software; however, they would require two distinct sets of lexicons :

.. a generic lexicon, of the present type, for translations which would be supplied as finished products to an end user

.. a series of subject-bound lexicons for translations to be supplied to a Translation Service as semi-finished products.

3.146 Comments on the results : revision rate

- The revision rate is significantly higher than in the first evaluation (5 %), which may partly explain the increase in the post-editing rates (texts which are more difficult even for human translators).

3.15 Summary table of the time required for manual translation and post-editing

3.1.1 Data

Doc. No	Number of SENTENCES - ORIG. TEXT	Number of WORDS - ORIG. TEXT	VERSION 3 MT (unrevised) Brussels	VERSION 4 MT (unrevised) Luxembourg	VERSION 4 MT (revised) Luxembourg	VERSION 5 PE Luxembourg	VERSION 6 PE Brussels	VERSION 7 PE DG XIII
1	37	929	-	1 h 50	2 h 15	2 h 00	-	0 h 40
2	16	291	1 h 30	2 h 10	2 h 40	1 h 20	1 h 35	0 h 25
3	45	983	-	2 h 35	3 h 35	1 h 55	-	0 h 57
4	18	281	1 h 35	2 h 40	3 h 00	1 h 05	1 h 05	0 h 28
5	36	860	-	2 h 12	2 h 42	2 h 00	-	1 h 01
6	19	301	1 h 40	1 h 30	1 h 50	1 h 15	2 h 55	0 h 26
7	65	951	-	2 h 02	3 h 02	2 h 30	-	0 h 59
8	19	256	1 h 25	1 h 23	1 h 53	1 h 00	1 h 30	0 h 22
9	46	858	-	1 h 45	2 h 20	3 h 00	-	1 h 02
10	20	345	1 h 20	2 h 35	3 h 05	1 h 45	1 h 15	0 h 17
11	64	922	3 h 35	1 h 50	2 h 15	3 h 30	0 h 45	0 h 57
12	18	308	1 h 30	2 h 20	2 h 45	1 h 30	1 h 15	1 h 01
13	42	879	-	2 h 10	3 h 20	2 h 30	-	1 h 16
14	16	232	0 h 45	0 h 55	1 h 40	0 h 45	1 h 00	0 h 27
15	38	877	2 h 35	2 h 55	3 h 15	2 h 10	2 h 30	0 h 57
16	16	385	0 h 45	2 h 30	3 h 05	1 h 40	2 h 00	-
17	57	975	3 h 05	4 h 40	5 h 40	1 h 50	8 h 30	-
18	19	319	1 h 00	3 h 00	3 h 30	2 h 30	2 h 30	-
19	39	790	2 h 00	4 h 30	5 h 15	1 h 10	1 h 00	-
20	26	561	-	2 h 15	2 h 45	0 h 45	-	-
TOTAL	656	12.303	22 h 45	47 h 47	59 h 52	36 h 20	27 h 50	11 h 15
TOTAL NUMBER OF WORDS V 1 IN FACT PROCESSED		-	6.282	12.303	12.303	12.303	6.282	9.273
AVERAGE TIME PER 100 WORDS OF THE NUMBER IN FACT PROCESSED		-	21.7 min	23.3 min	29.2 min	17.7 min	26.6 min	7.3 min
TOTAL V5/6	359	6.282	22 h 45	32 h 58	39 h 53	21 h 30	27 h 50	-
AVERAGE/ SENTENCE V 1	-	17.50	3.8 min	5.5 min	6.7 min	3.6 min	4.7 min	-
AVERAGE/ 100 WORDS V 1	-	-	22 min	31 min	38 min	21 min	27 min	-
TOTAL V5/7	499	9.273	-	30 h 52	39 h 37	28 h 25	-	11 h 15
AVERAGE/ SENTENCE V 1	-	18.58	-	3.7 min	4.8 min	3.4 min	-	1.4 min
AVERAGE/ 100 WORDS V 1	-	-	-	20 min	26 min	18 min	-	7 min
TOTAL V6/7	228	3.813	15 h 55	18 h 18	22 h 23	14 h 20	13 h 50	5 h 20
AVERAGE/ SENTENCE V 1	-	16.72	4.2 min	4.8 min	5.9 min	3.8 min	3.6 min	1.4 min
AVERAGE/ 100 WORDS V 1	-	-	25 min	29 min	35 min	23 min	22 min	8 min

3.152 Comments on the presentation of data

- As in the case of table in § 3.14, part totals have been calculated to allow comparisons.

3.153 Comments on the results

- For unrevised manual translation, staff of the Luxembourg Service take 46 % more time than their colleagues in Brussels; this can be explained by the fact that virtually all translations on the subject of agriculture are done in Brussels and the Luxembourg translators have to undertake a good deal of terminological research in this field.

- In the post-editing of machine translations, however, the Luxembourg Translation Service takes only 77 % of the time required by their colleagues in Brussels to do the same work (due to a lower post-editing : rate 35 % as against 48 %); their work can be considered of high quality since the final intelligibility achieved is equivalent to that of Brussels : 97.1 against 97.8 %.

- The DG XIII post-editor does his work in 7 to 8 minutes per 100 words (with 31 % post-editing), against :

  . 18 to 23 minutes for the post-editor of DG IX, Luxembourg (post-editing rate from 35 to 36 %)

  . 22 to 27 minutes for the post-editor of DG IX, Brussels (amount of post-editing : from 48 to 49 %)

  . 6 to 7 minutes for revision of the manual translation at DG IX, Luxembourg (amount of revision : 12 %)

  (the time brackets correspond to the various batches of documents assessed).

  This is obviously a rythm of work which could not be kept up in an official translation service with salaried translators, but it does give an idea (albeit by default) of the speed that could be achieved by a post-editor paid by the line. It is worth noting that despite this difference in speed, the result (intelligibility) falls into the same bracket as that of the other services.

3.16 Summary table of reading time (in minutes) (evaluators 1 - 7)

Version 1 : original text
Version 2 : machine translation
Version 3 : manual translation by DG IX, Brussels
Version 4 : manual translation by DG IX, Luxembourg
Version 5 : post-edited by DG IX, Luxembourg
Version 6 : post-edited by DG IX, Brussels
Version 7 : post-edited by DG XIII

3.161 Data

Doc. No	VERSION 1		VERSION 2		VERSION 3		VERSION 4		VERSION 5		VERSION 6		VERSION 7	
	NO OF SEN-TENCES	Min/ DOC.	NO OF SEN-TENCES	Min/ DOC.	NO OF SEN-TENCES	Min/ DOC.	NO OF SEN-TENCES	Min/ DOC.	NO OF SEN-TENCES	Min/ DOC.	NO OF SEN-TENCES	Min/ DOC.	NO OF SEN-TENCES	Min/ DOC.
1	37	10	37	24	36	20	36	24	34	18	36	32	37	37
2	16	05	16	08	16	10	16	08	16	05	16	06	16	09
3	45	09	45	22	45	15	45	17	47	13	45	17	45	18
4	18	11	18	04	18	05	18	05	18	12	18	04	18	03
5	36	27	36	15	36	13	32	10	35	15	36	11	36	08
6	19	08	19	08	19	05	19	10	19	08	19	06	19	03
7	65	11	65	22	70	14	65	10	69	10	70	23	65	11
8	19	03	19	07	19	03	16	04	19	05	21	07	19	03
9	46	13	46	19	47	05	46	13	46	20	47	16	47	09
10	20	06	20	04	19	06	20	04	20	04	20	05	20	08
11	64	13	64	12	64	16	62	05	64	20	64	05	(63	38)
12	18	03	18	07	21	07	17	05	17	08	18	10	(18	10)
13	42	12	42	19	45	11	42	10	42	07	45	11	(43	30)
14	16	04	16	09	16	02	16	03	16	03	16	07	(16	10)
15	38	05	38	22	38	09	38	08	38	08	38	20	(38	30)
16	16	10	16	11	13	05	16	02	16	05	16	03	-	-
17	57	10	57	18	58	14	56	11	58	12	58	07	-	-
18	19	10	19	08	19	03	19	03	18	04	19	04	-	-
19	39	20	39	15	41	10	39	12	39	07	39	08	-	-
20	26	12	26	15	26	08	26	05	26	08	26	07	-	-
TOTAL 20 DOCUMENTS	656	202	656	269	669	181	644	169	657	192	667	209	-	-
AV./ DOC.		10.10		13.45		9.05		8.45		9.6		10.45		-
V./SEN-TENCE	0.31		0.41		0.27		0.26		0.29		0.31		-	
NO OF WORDS	12,303		14,195		14,658		14,331		13,860		14,637		-	
AV./ 100 WORDS	1.64 min		1.89 min		1.23 min		1.18 min		1.38 min		1.43 min		-	
FIGURES FOR THE FIRST 10 DOCUMENTS IN THE SAMPLE														
MIN.	103		133		96		105		110		127		109	
NO OF WORDS	6,055		6,846		6,967		6,827		6,635		6,922		6,727	
AV./ 100 WORDS	1.70 min		1.94 min		1.38 min		1.54 min		1.66 min		1.83 min		1.60 min	

3.162 Comments of the presentation of data

- The figures for version 7 documents D 11 to 15 have not been taken into account: the post-edited versions of these 5 documents reached us the day before the group evaluation and we could only submit the manuscripts to the evaluators, whereas all the other texts were presented to them in a typed fair copy.

  This incident shows, however, the importance of a fair copy : the reading time for the first 10 documents of version 7 (fair copy) amounted to 0.34 minutes per sentence whereas the corresponding time for the 5 documents in manuscript form (D 11 to 15) was 0.66 minutes (almost double !).

- The reading times for the 20 documents in the sample, versions 1 to 6, cannot be compared with those for the first ten documents in version 7, even if such times are expressed as a ratio to the number of words. The influence of practice is clearly shown by a comparison between the reading times per 100 words for the first 10 documents and for the sample as a whole.

3.163 Comments on the results for the sample as a whole

- The reading time for the manual translation is 85 % of that for the original text, this being perfectly normal for evaluators with French rather than English as their mother tongue.

- The machine translation requires 57 % more reading time than the manual translation; this factor should not be overlooked in any economic assessment of machine translation.

- The post-editors of the translation services produce a text which reads slightly more slowly than the manual translation (on average 17 % more time required).

- The difference between the reading times for versions 5 and 7 and version 6 shows that a high rate of post-editing (48 % for V 6 against 36 % for V 5 and 31 % for V 7) does not improve the ease with which the text can be read : on the contrary, the first 10 documents of version 6 took 12 % more time to read than versions 5 and 7 of the same documents.

3.17 Summary table of the source of errors corrected during post-editing (V 5)

3.171 Data

DOC. NO.	No of words in machine translation (V2)	No of words post-edited (V5)	Amendments due to source text				Errors of acquisition	Amendments due to the translation system				Personal factors relating to the post-editor		Deviation
			Ambiguity	Incorrect or clumsy style	Syntactical error	Spelling		Lexicon	Analysis	Synthesis	Miscellaneous	Stylistic appreciation	Post-editing error	
1	994	355	-	27	1	1	6	126	22	18	2	154	-	+ 2
2	378	160	13	-	-	2	-	54	3	25	-	62	1	0
3	1111	391	-	-	-	-	12	104	50	28	19	178	-	0
4	362	132	-	-	-	-	-	42	22	19	-	52	-	+ 3
5	948	306	-	2	-	2	-	86	30	22	16	140	-	- 8
6	408	106	-	-	-	-	-	40	12	13	-	41	-	0
7	988	331	2	-	-	-	1	163	13	18	3	129	-	- 2
8	317	121	4	-	2	3	1	35	27	2	-	48	-	+ 1
9	934	346	21	-	13	11	1	97	39	16	11	131	1	- 5
10	406	147	2	-	-	5	-	81	13	2	2	41	-	- 1
11	976	377	24	-	4	10	14	106	46	25	5	150	-	+ 7
12	397	135	-	11	-	-	-	33	21	19	-	50	-	- 1
13	1043	423	1	8	-	4	1	148	55	27	11	161	2	- 5
14	277	142	2	-	5	4	8	77	8	9	1	25	-	- 3
15	1014	282	6	-	-	-	13	101	30	16	8	108	-	0
16			n o	e v a l u a t e d										
17	1178	411	14	-	-	2	2	166	36	14	14	163	1	+ 1
18	410	230	6	7	-	-	-	73	45	5	12	84	-	+ 2
19	923	238	3	5	-	-	5	75	31	40	13	64	-	- 2
20	622	199	2	-	-	-	-	76	23	22	4	64	-	- 3
TOTAL	13.686	4.832	100	65	25	44	64	1.683	526	340	121	1.845	5	- 14
%	-	100	2	1	1	1	1	35	11	7	3	38	-	-

3.172 Comments on the presentation of data

- Table 3.17 was drawn up on the basis of an analysis made by Mrs GUAZZO and incorporated into Table II of this report.

  Mrs GUAZZO carried out this work on the basis of :

  . version V 5 of the documents

  . the evaluation of error types made by evaluator E 8 of version V 5 which is given in § 3.14.

  The total number of corrections should therefore be the same in § 3.14 and 3.17. In fact, however, very slight discrepancies appeared and these have been recorded in the last column ("deviation") of § 3.17 : 14 corrections out of 4 832 words, i.e. less than 3 °/oo).

3.173 Comments on the results

- post-editing resulting from errors in the source text and errors in acquisition is insignificant : 6 % of the total corrections.

- a major part of post-editing (38 %) appears to be linked to the post-editor's assessment of the style of the translation.

  It seems, however, that this factor may have been over-estimated by the evaluator : in fact, if we compare the amounts of revision for version 5 documents (36 % correction by the post-editor of DG IX, Luxembourg) and 7 (31 % correction by the post-editor of DG XIII, whose aim it should be remembered was to minimize the post-editing), the amount of revision due to purely personal factors relating to the post-editor should be put at 5 %.

- however, there does not appear to be any dispute over which errors are due to the source text and which to acquisition.

- In view of this, the diagnosis can be adjusted as follows :

  . number of corrections due to source text (234) and acquisition (64) unchanged

  . number of corrections attributed to factors personal to the post-editor, reduced to 5 %, i.e. 242

  . breakdown of remaining corrections (1 850 - 242 = 1 608) among the following factors :

  * lexicon : $\frac{35 \times 1\,608}{56} = 1\,005$

  * analysis : $\frac{11 \times 1\,608}{56} = 316$

  * synthesis : $\frac{7 \times 1\,608}{56} = 201$

  * miscellaneous : $\frac{3 \times 1\,608}{56} = 86$

  which gives the following table :

3.174 Adjusted data

Source of errors	Gross number of		Adjusted number of	
	corrections	%	corrections	%
Source text	234	5	234	5
Acquisition	64	1	64	1
Lexicon	1 683	35	2 417	50
Analysis	526	11	756	16
Synthesis	340	7	487	10
Miscellaneous	121	3	184	4
Post-editorial assessment	1 850	38	676	14
Deviation	14	-	14	-
TOTAL	4 832	100	4 832	100

3.175 Comments on the adjusted results

- By far the most frequent reason for post-editing lies in the quality of the lexicon. Approximately half the corrections could be eliminated if the lexicons were accurate.

- This strongly suggests the need for a detailed study of possible strategies for the lexicon :

  . a single general lexicon, as at present

  . several specialized lexicons, as the system permits

  . pre-editing, that is recording of specialized mini-lexicons before the processing of certain documents or batches of documents on the same subject.

- We made comparison of the figures in the previous table with those in § 3.14 : in the latter, corrections are broken down according to a 'symptomatic' (in the medical sense) typology, whereas those in § 3.17 are broken down according to a 'diagnostic' typology.

  The relationships which might normally be assumed between symptoms and diagnosis have been indicated by arrows (solid line : strong relationship; dotted line : weak relationship)

Symptoms § 3.14		Possible relationships	Diagnosis § 3.17 (revised figures)	
Number of corrections	Factors		Factors	Number of corrections
2 638	Substitution	⟷	Lexicon	2 688
1 238	Correction	⟷	Analysis	842
416	New position	⟷	Synthesis	541
334	Deletion	⟷	Others (sources, acquisition, miscellaneous, assessment, deviation)	761
206	Addition	⟷		
4 832	TOTAL		TOTAL	4 832

A study of this table shows that, except for the link between 'correction by substitution' and 'lexicon', the other possible relationships are not borne borne out by reality.

- The conclusion to drawn from this is that there is no point in attempting to break down the total number of corrections (which is an essential criterion for the evaluation since it measures the amount of post-editing) into its various components (since it reveals nothing about the reasons for the errors)

- A breakdown of the corrections by source (§ 3.17), however, provides valuable information on the steps that should be taken to prevent such errors reoccuring. It proved, however, that such an evaluation could only be carried out by a specialist in the translation system and that it could produce very subjective results (particularly as regards the evaluation of the personal factor relating to the post-editor).

The figures given in § 3.18, which will now be examined, permit a different approach to be adopted on the same problem of diagnosing the causes of errors.

3.18 Analytical table of causes and remedies for the main errors in machine translation

3.181 List of abbreviations

LS : Limited semantics, noun expression
CLS : Conditional limited semantics : group of words translated otherwise than by a translation of each of the words in the group
HMRTN : Homograph routine
HM : Homograph
Dict. : Dictionary
STRPASS : 0 structural Pass...
" : 1 names of various 'SYSTRAN' analysis programmes
" 2 " " "
" : 3 " " "
STRPASS : 3 names of various 'SYSTRAN' analysis programmes
" : 4 " " "
v : verb
p.p. : past participle
pr.p. : present participle
inf. : infinitive
ADV : adverb
ADJ : adjective
PGM : programme

3.182 Data

DOC. No and sentence	Type - source of main errors	Possible remedies	Min. time (+)	Max. time (+)	Probable new sentence	Side effects	Comments
1/23	TERMINOLOGY 'Assumption' has been translated by 'hypothèse'; in the dictionary 'machinery' is 'machine' + ANALYSIS : bad word order : 'machinery' has been analysed in parallel with 'degree" instead of 'assumption'	Coding : expression 'assumption... by acceptation machinery = appareil ? Cannot be resolved since 'assumption' and 'machinery' are coded as abstract and concrete repectively and do not have a common semantic code	10m	30m	Acceptation par les.. .. et d'un appareil plus fort pour ...	none	
1/28	CODING : 'fill in' has not been understood as forming a single verb ovin is singular since it is only given as an adjective in the dictionary	Coding : fill.in (in one word) : combler mutton/lamb = ovin (always plural)	15m	30m	... doivent être comblés rapidement ... produits tels les ovins et les pommes de terre	none	
1/32	ANALYSIS : homograph, 'will' has been taken as a noun instead of a verb 'concerned' is taken as a conjugated verb and not as a past participle CODING : 'take account' is not recognized as a single expression	Programme : modify the homographic routine (=HMRTN) 003 (noun/conjugated verb/infinitive) taking into account a possible insertion modify HMRTN 020 (verb/past participle) preposition + noun + verb or participle = participle Coding : more refined CLS Take + account = tenir compte de	30m + 30m 5m	2 d 10m	... la Commission dans ses propositions tiendra compte de ... sur les marchés concernés	could be considerable possible	perhaps not worth the risk given rarity of expression
1/35	CODING : 'postponement' and 'up-to-date' are not in the dictionary TERMINOLOGY : 'include' is translated by 'comprendre' in the dictionary	Code postponement = renvoi up-to-date = récent, moderne amend include = inclure	6m 3m	15m	un tel renvoi permettra ... d'inclure des informations plus récentes dans le rapport ...	none possible	
1/37	ACQUISITION : bad punching of OJNO, etc. INPUT : No has no full stop	- -					'Number' is usually abbreviated No.

(+) m = minute
h = hour
d = day
wk = week

2/9	CODING : 'trickling filter' is not coded as an expression	LS trickling filter = filtre biologique	5m	10m	l'utilisation des filtres biologiques a donné des résultats semblables	none	
2/15	ACQUISITION : 'preformed' and not 'performed'!	-	-	-	-	-	only possible to correct by careful reading of the source text.
	TRANSFER : 'could' here is conditional and not imperfect	lexical routine	2wk	?		fair number in early stages	Included in next WTC improvement contract
3/19	CODING : "Slowing down" is not considered to be a noun	Code "ralentissement" as the noun form of the verb slow down, with HM 051	5m	10m	En 1977, le ralentissement ... (noté depuis 1975) a continué	none	
	HOMOGRAPH (Analysis) "Since" has been taken as a conjunction instead of a preposition	Programmes Modify HMRTN 061 preposition/conjunction/adverb	30m	?		fairly unlikely	
	"Noted" has been taken as a conjugated verb and not as a past participle	and HMRTN 020 verb/past participle (importance of incidental clause)	30m	2d		fairly unlikely here but always possible	included in next WTC improvement contract
3/34	CODING : increase has been given the "NARFL" code which makes it reflexive in French	Code delete this code	3m	?	... a augmenté plus lentement que...	none	
	SYNTHESIS : comparison of adverbs	Programme modify the synthesis PGM for adverbs	30m	4d		fairly unlikely	included in next WTC contract
4/3	ANALYSIS : Homograph error (Noun/present participle for "cooking" and consequently error in word order	Programmes revise HM032 (present participle/noun) and perhaps "STRPASS 2" the word order programme	60m ?	? ?	est préparé en cuisant, expulsant et dilatant du riz fin cru décortiqué ... contenant 75 % de protéines	grave risks	could be corrected by coding
	HOMOGRAPH : "fine" is an adjective here and not a noun	modify HMRTN 038 (noun/adjective)	30m	1d		always possible	included in next WTC improvement contract
	SYNTHESIS : treatment of %	improve synthesis	?	?		likely	
4/5	CODING : "Sweet course" is not coded as an expression	Code LS sweet course = dessert	3m	10m		none	
	TERMINOLOGY : "covered" is coded "couvert" in the dictionary	Need not be changed			..... l'épaississement des desserts contenant du lait		covered is most frequently translated by couvert
	SYNTHESIS : definite article in front of lait	Programme modify article PGMs in the synthesis	?	?		likely	very difficult
4/8	ANALYSIS : HM "pulp" here is a noun and not a verb	Virtually impossible for HMRTN to determine here			-		not worth changing

5/9	ANALYSIS : "To be necessary" is not the complement of "believes" because it has not been recognized that the relative THAT is missing hence : Series has not been recognized as the subject of is given	Programme modify STRPASSO, clause separation	2h	7d	la série d'améliorations que la Commission croit être nécessaire (s) est donnée(s) dans sa totalité au-dessous	very likely	difficult to resolve
	CODING : "in full" has not been recognized as an expression	Code CLS IN + FULL (noun) = dans sa totalité	10m	-			
5/11	CODING : Bull. is not in the dictionary EC is not in the dictionary	Code Bull. = Bull EC = CE	7m	5m + ?	Bull. CE	none possible	abbreviations in English have many different meanings
5/18	CODING : "Across-the-board is not in the dictionary as an expression	Code across-the-board adjective = à usage multiple ou d'emploi général			les instruments d'emploi général		
	ANALYSIS : HM "markets" is a noun and not a verb here or principal is an adjective and not a noun	Programme improve HMRTN 003 noun/verb/infinitive preposition + adjective + noun/verb = noun preferable to improve HMRTN 025 (adjective/noun) preposition + noun/adjective + verb/noun = adjective !	15m	1d	.... des principaux marchés	fairly unlikely fairly unlikely	included in next WTC improvement contract it would also be possible to code ILS Principal market
5/19	CODING : cf. 18						
5/23	CODING : cf. 11						
5/24	TERMINOLOGY : "relate" is translated by rapporter	Code CLS ? Relate + object + To = mettre en rapport	10m	-	de plus en plus mis en rapport ... doit viser à fournir un revenu comparable à ce	none	this is a false sentence artificially separated by : " : "
	CODING : "aim" + infinitive = viser à	AIM = viser + à + infinitive	3m			none	
	"earned income" is an expression	CLS = revenu (without gagné !)	5m		la productivité et.. ...	none	major problem for pre-editing ?
	TERMINOLOGY : the pronoun that is translated "ce" in the dictionary	lexical routine ?	?	?		highly likely	very tricky ! and difficult
	SYNTHESIS : indefinite article before productivity	Coding : give an article code (definite or nothing) to productivity	2m	-			
6/5	TERMINOLOGY : cultivar "dunes"	Code or programme by developing a routine for proper nouns	3m	10m	cultivars Dunes ... ont donné les productions de saison sèche les plus élevées et ...	none probably none	unusual term (there will always be work of this type in any text)
	ANALYSIS : "dry" is not been recognized as an adjective describing "season"	Coding : LS : dry season	5m	10m		none	
6/13	TERMINOLOGY : "clam" has been translated by peigne	Code Clam = palourde ?	3m	-	extraits ... de palourdes ...	none	peigne is also a French noun for "clam"

6/19	ANALYSIS : Homograph : "further" has been taken as an adverb. Hence, studies has been taken as a verb	modify HMRTN 057 (adj/adv.	1h	7d	on suggère qu'une étude ultérieure ...	probably none	Part of next WIC contract
	SYNTHESIS : "should" has not been translated	PGM : Analysis - Write a routine for auxiliary without infinitive	?				Problem of ellipsis. Very tricky !
	HOMOGRAPH : "shelling" has been taken as a noun instead of the present participle	modify HMRTN 051	30m	?	de décortiquer et de nettoyer ...	possible	
7/26	CODING : "one or more" constitutes an expression	Code one.or.more as an Idiom = un ou plusieurs treated as an article	5m	10m	un ou plusieurs projets à ...	none	
	Semi-public is not in the dictionary				qui peut être publique, semi-publique ou provée		Unknown words will always pose problems, even of analysis
	SYNTHESIS : agreement of adjectives or more likely due to semi-public : unknown word (hence homograph error of Public)	Code semi-public = semi-public				none	
7/34	ARTICLE : No article required before catégorie	Coding : CLS By + category = par catégorie	5m		à inclure par catégorie de propriété	none	
	TERMINOLOGY : "include" has been translated by comprendre	it could be changed to inclure ?	3m				
	"land ownership"	LS = propriété (quite simpl	5m			none	
7/37	CODING : of general concern to : translated by à	Code concern = intérêt to = pour	3m	15m	d'intérêt général pour ...	none	
	TERMINOLOGY : Catchment area	LS : bassin de captation	10m	-	un important bassin de captation	none	
7/38	CODING : "mitigate" is not in the lexicon	Code = amoindrir	5m	?	des mesures pour amoindrir	none	
7/40	CODING : "efficacity" is not in the lexicon	Code = efficacité	3m		en ce qui concerne leur efficacité	none	Not worth is at the present time
	TRANSFER : "as to" here should be : en ce qui concerne, or quant à ...	lexical routine	7d	?		likely	
7/41	Infinitive instead of future because of break in sentence	difficult					
7/56	HOMOGRAPH : "before" has been taken as a conjunction and not as a preposition	PGM modify HMRTN 061 (adverb/conjunction)	30m	-	avant l'échéance de la ... visée au § 1 le conseil examine ce règlement sur ...	few but possible	the problem could perhaps have been avoided by separating the "2", in the sentence by punching (refine coding rules)
	CODING : in = en	Code	3m	-			
8/2	CODING : "spray irrigation" is an expression	Code LS : irrigation par vaporisation aspersion	10m	20m	nouvelles méthodes d'irrigation par vaporisation	none	
8/3	INPUT : en for in						

8/4	HOMOGRAPH : "later" should be an adverb "later on" is an expression; hence "reviews" has been taken as an noun	PGM modify HMRTN 057 or code CLS later + on ? = plus tard or par la suite			plus tard, cette érude détaillée étudie ...	possible	make coding changes first
8/7	HOMOGRAPH : "increased" has been taken as a conjugated verb	due to break in sentence (:) probably improve division of sentences	?	14d			there are many cases of this type
8/16	CODING : "nitrogen fertilization" is an expression TERMINOLOGY : "field" is "domaine" in the lexicon	Code : LS : fertilisation à l'azote by lexical routine or coding	7m 1d ?	10m 10m	on a étudié la fertilisation de l'azote dans les pots et en pleine terre	none fairly unlikely	
8/19	HOMOGRAPH : "that" has been taken as a demonstrative adjective instead of a conjunction	Style : give article to amount changes could be made to HMRTN 045 on that	- 60m	-	montrent que la quantité	?	there is a limit to possible analysis of telegraphic style
9/14	ANALYSIS : antecedent of pronoun "one" HOMOGRAPH : "concerned" and "used" are taken as conjugated verbs	lexical routine or analysis of pronouns change HMRTN 020	1d 1d ?	1d 1d 7d	était celle ... dans le pays concerné, l'année employée par...	possible possible	included in next WTC contract Difficult to resolve
9/19	INPUT : Hols HOMOGRAPH : "groups" is taken as a noun	eliminate errors in source text modify HMRTN 025 (noun/verb)	 60m	- -	 dans le groupe d'exploitation groupe 10 ou plus ..	- fairly unlikely	
9/20	TERMINOLOGY : the abbreviation is not in the dictionary	impossible to think of all possible abbreviations !					abbreviations of 2 or 3 letters appearing alone may have 100 different meanings
9/25	cf 20						
9/27	CODING : "at most" has not been recognized as an expression	CLS coding or lexical routine	- 60m	15m -	une unité de travail au maximum	none	
9/34	INPUT : captial CODING : "book value" is an expression Après le commancement badly coded	eliminate spelling errors Code LS : valeur comptable modify the proposition at + beginning	 8m 	 10m 	 est la moyenne des valeurs comptables au début ...		
9/36	PUNCHING : a space was required between 30 and)	punching in accordance with rules					punching rules should be simplified
9/42	INPUT : assigend	correct spelling mistakes					
9/44	HOMOGRAPH : "forming" has been taken as a noun "round" has been taken as a past participle "harvests" has been taken as a noun	Coding Form + part = faire part PGM modify HMRTN 023 past/participle/noun) and HMRTN 025 (verb/noun)	10m 60m 60m	10m 2d 	faisant partie qui restent dans le sol ...	perhaps always possible	this coding would have to be tested included in next contract

9/46	INPUT : Byt	correct spelling mistakes					
10/1 2	CODING : of expressions	Code LS s	5m 5m	10m 10m	culture de plantes stimulantes - plantes médicinales	none	
10/7	CODING : "cutting down operation" is an expression	Code cLS, opération de coupe	15m	20m	les opérations de coupes peuvent être effectuées au cours des mois de	none ?	
	TRANSFER : en les mois	extend lexical routine on months	30m	-		none	
10/10	TERMINOLOGY : pas d'erreur ...	impossible to correct					natural ambiguity
10/11	CODING : "premature flowering" is an expression	Code LS floraison précoce	7m	10m	à l'étude de la floraison précoce de	none	
	"Hévéa" was not in the dictionary		3m		l'hévéa	none	
10/12	CODING : "bring into flower" is an expression	Code CLS amener à la floraison or faire fleurir	15m	30m	méthode pour faire fleurir de jeunes ...	none	might not work
10/13	TERMINOLOGY : "month old" and "plant" = installation in the dictionary	Coding or lexical routine Coding of CLS ?	10m 150m 10m	10m ? 10m	25 % de plantes de 16 mois	none	unlikely that this context will arise again
10/14	HOMOGRAPH : "plant" is taken as an adjective	modify HMRTN 038 (noun/adjective) or code LS	30m 5m			fairly unlikely	difficult context
	TERMINOLOGY : "ground level" is translated "au niveau du rez-de-chaussée"	probably by coding		10m	au niveau du sol... le pourcentage de plantes cassées	possible	
10/16	Break in sentence makes interpretation of the present participle difficult	Change punching rules	?			possible	
10/18	Source text is ambiguous	impossible					
10/20	INPUT : "later" for "latter"	correct spelling mistake					
	TERMINOLOGY : Towards is translated by "vers"	Code CLS tolérance towards = envers	7m	-	la tolérance envers le traitement...		
11/2	CODING : "in the chair" is an expression	Code CLS à la Présidence	10m	10m	à la Présidence ...	fairly unlikely	
11/16	CODING : "expert group" is an expression	Code LS	5m	10m	des groupes d'experts furent chargés du mandat ...	none	
	"entrust" and "mandate" are not in the lexicon	Code 2 - charger de - mandat	3m 3m	10m 10m			
	INPUT : "restreints"	correct spelling mistakes	-	-	-		
11/25	CODING : "fortnight" not in the dictionary	Code quinzaine de jours	3m	10m	on a demandé aux délégations ... dans une quinzaine de jours...	none very likely	minor error
	SYNTHESIS : English passive = on	PGM modigy synthesis	90m	?			
11/41	CODING : expert group advise ... of	cf. 16 code CLS conseiller X pour			conseiller le CPRA pour l'établissement	possible	
11/54	INPUT : "meaning" instead of "weaning"	Correct spelling mistakes					
11/57	INPUT : "non" instead of "not"	Correct spelling listakes					

11/59	INPUT : break in sentence					possible	new punching rules
	ANALYSIS : Homograph "that" is a relative and not a conjunction ⇒ "meant" is a past participle and not a verb ⇒ "that" is a pronoun and not a conjunction (ambiguity in source text)?	PGM modify HMRTN 045 on that	90m	2d	a été informé que les projets ... signifient que ...	possible	included in next WTC improvement contract
11/63	TRANSFER : for the future : = pour	Code IDiom - à l'avenir	10m	-	à l'avenir ...	none	
	CODING : "Third country" is an expression	Code LS pays tiers	5m		permettre à un pays tiers		
	"COST" has been taken as a noun and not as an abbreviation					immense	impossible to allow for all abbreviations
11/64	INPUT : "hereby" instead of whereby	avoid spelling mistakes					
	ANALYSIS : "Community" qualifies projects and not "interest"	Code IDiom Community interest	3m	-	aux projets d'intérêt communautaire	none	
12/6	CODING : "Saveloys" is not in the dictionary	Code = cervelas	3m	-	..., cervelas),	none	
	"culture media" is an expression	Code LS bouillon de culture	10m	-	en utilisant les 4 bouillons de culture suivants ...	none	
12/15	Source text						Clumsy, but reflects English text.
12/18	HOMOGRAPH : "official" has been taken as a noun and not as an adjective	PGM modify HMRTN 038 (noun/adjective)	30m	2d	première actions officielle	fairly unlikely	included in next WTC contract
13/8	CODING : gouvernemental incorrect article for disponibilité	Code modify synthesis code for gouvernemental	5m		influence gouvernementale ... la disponibilité ...	none	
		Give article code to disponibilité	2m				
13/19	CODING : Inaccurate coding of "agricultural production"	Code modify LS	-	15m	les entreprises ... par une production agricole ... de transformation	not worth	not worth it
	"production-processing" is a composite word which is not in the lexicon	code ?	5m	-			
	TERMINOLOGY : "undertaking" is translated by "engagement" in the lexicon	impossible to change the meaning to "entreprise"					
13/23	CODING "ever-decreasing" is not in the dictionary as a single word	Code toujours décroissant	5m	10m		none	
	"agricultural undertaking" is an expression	exploitations agricoles	5m	10m		none	
	ANALYSIS : "process" has been analysed in parallel with 'demand' instead of 'increase'	PGM modify STRPASS 2 (not only ... but also)	60m		... par un nombre toujours décroissant ... mais aussi un processus de sélection parmi des exploitations agricoles ...	none	

13/26	HOMOGRAPH : 'This has been taken as a pronoun	PGM modify HEARTN 046 (pronoun adjective/pronoun)	30m		ce facteur ... a été accentué plus encore par ...	none	
	CODING : agric. undertaking	cf. 23 code					
	"everincreasingly" is not in the lexicon	sans arrêt ?	10m		des exploitations agricoles fonctionnant (plus) efficacement	none	rare
	SYNTHESIS : position of adverbs	modify re-arrangement	10m			probably none	
	ANALYSIS : "operating" has been taken as a gerund instead of an adjective	modify STRPASS 1	120m			very likely	tricky
13/32	CODING : "on the X-side" is an expression	Code CLS du côté de X	10m		du côté de l'approvisionnement ... et par la stratégie et la force du marché des compétiteurs	none	
	INPUT : strategy	avoid spelling mistakes					
	TERMINOLOGY : competition is "concurrent"	code CLS	1m	10m			
13/33	CODING : "some-of-the" is not in the dictionary	Code certain	5m	10m	certains facteurs ... côté des demandes ...	none	due to a punching error, 'some.of. the' was rejected during an updating
	on the side ...	cf. 32					
	INPUT : High-purchasing-power	code LS 'pouvoir d'achat'	5m		de la demande d'un pouvoir d'achat élevé	none	
13/34	CODING : "decision-making" is a set expression	Code decision	10m	10m	le processus de décision à savoir si oui ou non ...	none	
	"concerning"	Code CLS à savoir si oui ou non or lexical routine	15m 1d			possible	present participles do not usually gover conjunctions
13/35	CODING : "struggle" and "venture" are not in the dictionary	Code lutter + lutte risquer	7m 5m	10m 10m	de survivre au processus de sélection provenant de la théorie et la lutte de concurrence ... par risque pour appliquer des méthodes	none none	
	"survivre" + object takes the preposition à	modify survivre by CLS	5m			none	
	CODING : arise + out of	Code CLS provenir ... de	10m			none	
13/36	TERMINOLOGY : "lead" is translated by "plomb" in the lexicon	Code by LS or CLS	10m	15m	et la productivité principale	possible but fairly unlikely	
13/41	CODING : "take over" is one verb -	Code IDiom or CLS : prendre sous contrôle	10m	15m	ont été prises sous contrôle par des non-agriculteurs	fairly unlikely	
	gouvernemental	cf. above					
	TERMINOLOGY : "undertaking" is translated by "engagement" in the dictionary	Code LS exploitation agricole	10m				

14/3	CODING : "Please" is not in the dictionary as an adverb	Code S.V.P. ?	5m			possible because of HM	
	ANALYSIS : The imperative form has not been recognized	PGM developed an analysis of imperative forms	?	30d	envoyer s.v.p. ...	certain	difficult task, should only be undertaken if this type of document is actually going to be translated by AT.
14/7	HOMOGRAPH : "what" is interrogative here or analysis of interrogation structures	improve analysis of interrogative forms	?	30d	quels sont ...	very likely	part of next contract with WTC
	CODING : "end-of-lay" is not in the lexicon	rare					not worth the trouble
15/1	TERMINOLOGY : acetylene inhibition	Code LS inhibition par acétylène	5m		utilisation de l'inhibition par acétylène de protoxyde d'azote	none	
	CODING : nitrous oxide	Code LS protoxide d'azote	5m	10m		none	
15/3	HOMOGRAPH : "cycles" has been taken as a verb	Code LS would be adequate or PGM modify HMRTN 025	5m	10m	du cyle de l'azote.. .. la réduction de nitrate ou de nitrite à	none fairly unlikely	
	ANALYSIS : "nitrite" has been analysed in parallel with "reduction" instead of "nitrate"	PGM modify STRPASS 2 taking into account semantic categories					
15/7	energy-intensive	avoid punching errors			sur l'industrie des engrais de l'énergie intensive		
15/11	CODING : nitrous oxide "denitrify" is not in the lexicon	cf. 1 Code dénitrifier	5m		la réduction du protoxyde d'azote a été bloquée évidemment, comme confirmé plus tard ...	none	
	TRANSFER : "as" here is "comme" HM error of "later" and "confirmed"	PGM develop lexical routine for as	2h	2 - 3d		possible	part of next contract with WTC
15/25	CODING : WHC is not in the lexicon	-				cannot allow for all possible abbreviations in English	
15/26	Division of sentence						
15/30	HOMOGRAPH : "tried" has been taken as a verb not as a past participle	PGM modify HMRTN 020	60m	7d	sol séché ...	possible	part of next contract with WTC
15/35	cf. 25						
15/36	ANALYSIS : Because of 5 abbreviations	impossible					5 abbreviations (which are not in the dictionary) make analysis extremely difficult
15/37	TRANSFER : "HE" here is the gas	by coding of a CLS ?	15m		était soit l'He soit l'air ambiant	possible	

16/2	CODING : "abet" is not in the lexicon	Code accentuer	5m	10m	accentuée par des ruptures dues au temps dans l'écoulement ... ont poussé plus haut les prix des aliments au détail pendant l'hiver	none	
	"weather-related" is not in the lexicon	Code dû au temps	10m			none	impossible to code all terms made up of several words
	HOMOGRAPH : "Food" here is a noun and not an adjective	Code LS retail food price ?		15m		possible	might not work
16/3	HOMOGRAPH : "Further" is an adjective and not an adverb	PGM modify lexical routine	30m	1d	d'autres accroissements sont probables par la moitié de l'année	fairly unlikely	part of next contract with WTC
	CODING : "midyear" is not in the dictionary	Code = la moitié de l'année					
16/5	CODING : "Retail food price" is an expression	cf. 2					
	INPUT : bad English	better style					"likely" is not an adverb. Should be "will probably average"
16/6	CODING : "first quarter" is an expression	Code premier trimestre	5m	10m	toute l'augmentation du premier trimestre est associée des produits alimentaires produits aux USA	fairly unlikely	
	"US-produced" is not in the lexicon	Code produit aux USA	5m	10m			strange word
16/7	CODING : "last fall" is an expression	Code LS automne dernier or	5m		... pendant cet hiver que les bas saisonniers l'automne dernier	probable	this is an Americanism and only worth taking into consideration if American texts are to be translated
	"fall" is in the dictionary = chute	lexical routine	3h	1d			
16/8	CODING : last fall	cf. 7					
	INPUT : bad English	better style					cf. 5
16/9	CODING : "first quarter at home and away-from home"	cf. 6 Code impossible to think of personal stylistics !					strange terms
16/10	CODING : first quarter	cf. 6 and 9					American usage
	through	better style					
	HOMOGRAPH : "Half" is taken as an adverb instead of a noun	PGM improved HMRTN 052			... la deuxième moitié	possible	
16/11	CODING : US-produced	cf. 6					
	INPUT : bad style	better style					cf. 5
16/12	TRANSFER : "expect" is in the lexicon : s'attendre à	PGM develop the lexical routine	5h	2d	on s'attend à ce que les prix ...	possible	part of next contract with WTC
	CODING : "By Spring" is an expression	Code Idiom d'ici le printemps	5m	10m	... d'ici le printemps	fairly unlikely	
16/13	CODING : "second quarter" is an expression	Code LS deuxième trimestre	5m			probably none	
	bad English	better style					cf. 5

16/14	CODING : "Spring" is translated by "ressort" in the dictionary	Code CLS le printemps dernier or PGM lexical routine in spring	5m 5h	10m		possible possible initially	
16/15	HOMOGRAPH : "Price" here is a noun and not a verb (due to "further") INPUT : bad English	PGM modify routine for further (cf. 3) better style	30m 	1d	d'autres baisses des prix ("about" is never used as an adverb before a verb - should read : "approximately" - probably bad Americanism)	fairly unlikely	part of next contract with WTC
16/16	TRANSFER : expect CODING : "upward pressure" is an expression "overall food price" is an expression U.S. is also an adjective	 cf. 2 Code LS montée Code LS prix des aliments en gros Code américain	 5m 8m 8m	 10m	on s'attend à une montée des aliments en gros des produits alimentaires produits aux exploitations américaines	 possible none possible because of HM	
17/7	CODING : "gram negative" is an expression "nonspore forming" is an expression	Code LS gram-négative LS non sporifère	5m 10m	10m	la tige non sporifère gram-négative la plus significative liée ...	none	
17/9	ANALYSIS : "family enteros..." is not in apposition	Code family as STRAP strong apposition requirer and modify PGM STRPASS 1 and rearrangement	 3m 5d		membres de la famille entéro-bactériacée (famille des entérobacté ...)	 likely	time-consuming and expensive work and changes would also have to be made in the dictionary
17/30	CODING : Vacuum packaged	Code CLS emballé sous vide	8m	10m	échantillons emballés sous vide	none	error also due to break in sentence
17/31	HOMOGRAPH : "develop", "sealed" have been taken as verbs Proper noun CODING : "mil" is not in the dictionary "vacuum packaging" is an expression	PGM modify HMRTN 020 (verb/past participle) design special processing Code mil = ml Code LS emballage sous vide	90m 15d 5m 8m	2d 10m	 poche de polyéthylène de 1.5 ml (All-Vak 13, Intern Kenfield Distribution Co, Broadview, Illinois) avec le vide de guage développé de 25 livres par pouce carré et soudé à chaud ...	fairly unlikely none none	part of next contract with WTC usually written ml
17/33	CODING : "mil"	cf. 31					
17/35	CODING : "mil" HOMOGRAPH : "folded" is a past participle here "fold over" is one verb	cf. 31 and 33 PGM modify HMRTN 020 (verb/past participle) Code idiom or CLS replier	 90m 10m		 des sacs en polyéthylène de 1.5 ml avec le dessus des poches replié ...	 possible none	
17/40	CODING : "raw broiler part" is an expression Proper nouns	Code partie à rôtir cf. 31	15m	15m	... des parties crues à rôtir ... (Plastique Falcon Los Angeles, Calif.)		
17/45	CODING : "Plant count agar" is an expression	Code LS	10m	10m	gélose de compte (DIFCO)	none	

Note

In the case of sentences D 5/10
D 6/14
D 9/39
D 17/34,

which were assessed as having less than 50 % intelligibility by the evaluator (E 9), Mrs GUAZZO and Mr PIGOTT did not analyse the errors since the unintelligibility of these sentences is not due to a bad translation but to the wording in the source language.

3.183 Comments

- Table 3.18 was drawn up by Mrs GUAZZO, with the assistance of Mr PIGOTT; we are grateful to them for having agreed to carry out this valuable if time-consuming exercise

- In our view, this analysis is one of the key factors in the evaluation, since it makes it possible to assess :

  . to what degree the system can be perfected
  . the time required to incorporate these improvements
  . the possible side effects which such amendments might have on the translation of other sentences

- From this evaluation, there emerge two important points which are quoted by Mr PIGOTT :

  "Altough we have both nearly always been able to identify an error and explain how it could be corrected, it is not necessarily the case that it would normally be corrected in this way.
  Indeed, many of the errors which occur result from uncommon structures in the source language. However, most coding and program changes are only undertaken if the same type of structure is likely to occur again.

  It is interesting to note that about half the programming errors identified have already been recognized as belonging to the error types to be handled in the next improvement contract.
  I would therefore expect a considerable measure of translation quality improvement to be achieved over the next 12 months, particularly if a continuing effort is exerted on dictionary improvement."

- Table 3.182 gives an indication of the methodology which will be adopted in the future for improving the system : so far two different teams have been responsible for this work :

  . that of Mr PIGOTT, in Luxembourg, which was concerned with :
    .. amendments to the dictionaries; these changes being inputted to the computer in batches at certain intervals : up-dating of the dictionaries every two to four weeks
    .. the types of change to be made to the programme itself; these changes were grouped together and covered by an overall contract with WTC

  . that of WTC, in La Jolla, which carried out the changes required in the translation programs on the basis of contracts drawn up by the Luxembourg team.

In future, there will be a single team in Luxembourg, under the Kranklin Institute, which, for language pairs other then English-French, will receive the machine translations, analyse them with the help of a scale of ratings probably similar to that of Table 3.182 and, on the basis of this analysis, decide what changes should be made in both the dictionaries and the translation programs.

- As it can be seen from Table 3.182, the latter is primarily a tool for making a qualitative analysis which only provides a list of possible corrections without giving information on the frequency with which such errors occur and hence it does not provide a basis for decision-making allowing one to select, in a less empirical fashion than at present :

  . the amendments to dictionaries and programs which should be given priority

  . amendments which need not be taken into account because the corresponding errors arise from structures rarely found in the source language (cf. previous page, comment made by Mr PIGOTT).

  Unfortunately it appears that the incorporation of quantitative factors into the table is a difficult proposition : to do so one would require not a sample of a few hundred sentences as in the present evaluation but several tens of thousands sentences and, for each of them :

  . the original text
  . the machine translation
  . the post-edited version (preferably consistent for each error in different sentences !)
  . an unequivocal identification of corrections on the basis of which they could be put into groups and their frequency calculated as well as an evaluation of the consequences (side effects) of any change.

  It appears difficult to put such a method into practice since it would require long and no doubt expensive preparatory work; it seems, however, that it might be worth studying in greater detail and trying, if not to replace the empirical methods currently used, at least to complement them by a more systematic approach.

3.19 Acceptability

3.191 Analysis of the 17 questionaires

	Institutions		
	CEC	CNRA	Gembloux
Number of questionnaires returned	4	10	3
Basic training (more than one diploma possible)			
- Agronomy (doctorate, degree in engineering)	1	6	3
- Sciences (doctorate in science, doctorate in organic chemistry, degree in chemical engineering, degree in natural sciences)	-	3	1
- Engineer	-	1	-
- Laywer	2	-	-
- Degree in commerce and finance	1		
Total	4	10	4
Type of work			
- Research - plant life	-	3	-
- Research - foodstuffs	-	1	1
- Research - scientifico-economic field	-	2	-
- Research (field not specified)	-	2	-
- Management, administration	4	-	-
- Documentation (for research scientists and practitioners)	-	2	2
Specific fields of activity			
Fields of activity			
- Improvement of leeks and onions	-	1	-
- Animal foodstuffs	-	1	-
- Regional socio-economic analysis	-	1	-
- Environment, energy, systems, innovation	-	1	-
- Microbiology	-	1	-
- Study of the quality of vegetable products under various forms of conservation	-	1	-
- Soil science	-	1	-
- Photosynthesis, pigments, enzymatic activity	-	1	-
- Scientific documentation, agronomy	-	1	2
- Documentation for practitioners (technical and economic)	-	1	-
- Tobacco, hops, seeds, potatoes	1	-	-
- Agricultural law	2	-	-
- Agricultural economics	1	-	-
Languages			
Read Danish	-	-	-
Dutch	2	-	-
English	3	10	3
French	4	10	3
German	3	6	-
Italian	3	4	-

		CEC	CNRA	Gembloux
Written	Danish	-	-	-
	Dutch	2	-	-
	English	3	8	3
	French	4	10	3
	German	3	2	-
	Italian	1	-	-
Does your work involve the subject dealt with in at least one of the documents in the sample which you have examined ?				
	often	4	6	3
	sometimes	-	4	-
	never	-	-	-
Do you consider that the translation of these documents is acceptable, knowing that they are produced by a computer and can be obtained in a very short period of around half a day ?				
	in all cases	1	1	-
	in certain circumstances	3	7	3
	never	-	2	-
	for myself	4	5	-
	for certain of my colleagues and assistants	3	8	1
In what circumstances do you consider this translation acceptable ? (various replies possible)				
-	Reader does not know source languages	-	3	3
-	Reader familiar with subject matter	1	2	-
-	Speed	1	1	-
-	Availability of source text	-	1	1
-	Current awareness	1	1	-
-	Source document written in simple style	1	-	-
-	To assist a translator	-	-	1
-	No answer (because results considered unacceptable)	-	2	-
-	No answer (although results considered acceptable)	2	1	-
In what cases might this quality of translation be adequate for your purposes ? (various replies possible)				
-	To assess whether a document is worth reading and to have a general idea of the subject	-	3	-
-	Current awareness	1	4	-
-	Very low cost	-	2	-
-	For the translation of internal working documents	2	-	-
-	Language not known	-	1	-

	CEC	CNRA	Gembloux
- Familiar subject	-	1	-
- Availability of original text	-	1	-
- In the vast majority of cases	-	1	-
- In no circumstances	-	1	1
- No answer (because results considered unacceptable)	-	2	-
- No answer (although results considered acceptable)	1	1	2
In what cases might this quality of translation be useful to you, on condition that you could subsequently request a manual translation or a version of the machine translation corrected by a human translator ? (various replies possible)			
- Languages not known	-	6	-
- For important texts (basic articles, most important working documents)	2	1	-
- Subject with which the reader is not familiar	-	1	-
- In all cases	1	-	-
- No answer (because results considered unacceptable)	-	2	-
- No answer (although results considered acceptable)	2	-	3
Which of the defects of the enclosed machine translations seem to you the most inconvenient ? (various replies possible)			
- Lexicography			
. Words not translated	3	2	1
. Words badly translated (homonyms)	1	3	1
. Expressions badly translated	-	-	1
. Inadequate technical terminology	1	-	-
- Pre-editing			
. Translation of proper nouns (names of persons, companies...) which should not be translated	-	3	1
. Incorrect translation of symbols and abbreviations	-	3	3
. Distorsion of formulae	-	2	-
. Translation of acronyms which should not be translated	-	1	-
- Syntax			
. Incorrect agreement of adjectives in genitive constructions or enumerations	-	1	2
. Incorrect rendering of tenses in the conjugation of verbs	-	-	3
. Incorrect translation of inversion in English	-	1	-
. Turns of phrase	1	-	-

	CEC	CNRA	Gem-bloux
- Misinterpretation and wrong meaning	1	5	-
- Incomprehensible sentences	-	1	-
Total	7	22	12
Would you be interested in having access to an automatic translation system providing texts of the quality of those submitted to you ? In what direction(s) ?			
- From your mother tongue into another language	1	1	-
- From another language into your mother tongue	2	6	1
- Yes, without indication of the direction	1	-	-
- Yes, for more obscure languages than English	-	1	-
- Yes, if the quality were better	-	-	1
- No	1	1	1
- No answer	-	2	-
If yes, what benefits would you gain from such a system ?			
- It would save time	3	3	-
- Quick assessment of the level and interest of a document in an unknown language	1	3	1
- Much better dissemination of our articles in English-speaking countries	-	1	-
- No answer	-	3	2
Personal comments made by those taking part in the survey			
- This project is an extremely interesting experiment; however, at the present stage of its development it cannot be of great use ... particularly since it is designed only for translation from English ..... . In my view, one's judgement would be less severe in the case of a language less well-known than English	-	1	-
- The mistakes made by the computer are rare or simply ones of form	-	1	-
- The errors of syntax or spelling do not represent a major defect insofar as the reader has been warned to expect them	-	1	-
- The level of machine translation achieved these year cannot serve as a basis for work or decision	-	-	1
- In the case of legal documents, revision would take longer than manual translation	2	-	-
- The results of the machine translation are poor when the sentence in the source text is too 'elegant' or too long	1	-	-

	CEC	CNRA	Gem-bloux
- Words not translated should stand out in the target text (for example between inverted commas)	1	-	-
- Machine translation is only worthwhile if its cost is minimal	-	1	-
- Machine translation must be restricted to one-off requests; the printing and widespread circulation of texts translated automatically should not be envisaged	-	1	-

3.192 Comments on the scope of the survey

The questionnaire was completed by a limited number of potential users of machine translation (17 persons), and in a limited number of organizations (3 institutions), and only covered a single pair of languages, English-French, in one field, that of agriculture and the food industry.

Our action should therefore be regarded as a preliminary survey and not as an exercise in market research in the true sense of the word.

The results must therefore be interpreted merely as trends.

It should also be noted that the 17 replies were all given by persons actually concerned with the field covered by the evaluation (see replies to the question 'Does your work involve the subject dealt with in at least one of the documents in the sample').

3.193 Comments on the results

16 of the 17 participants have a reading knowledge of English. A survey on a translation system with a source language other than English would probably have revealed more about user requirements.

It is clear that machine translation at its present level does provide a useful service on condition that translation can be obtained within a very short period (of the order of half a day) :

- 15 positive replies against 2 negative replies to the question 'Do you consider the translation of these documents to be acceptable ...'

- 12 positive replies and 5 negative replies to the question 'Would you be interested in having access to (such) a system ...'.

Machine translation would appear to be particularly appreciated for translation :

- from less well-known languages
- into the mother tongue of the user
- working documents where it is necessary to get a rapid idea of the contents (CEC)
- scientific and technical documents when the reader wishes to assess their relevance and get some idea of the contents (CNRA and Gembloux).

A machine translation system will only be used if it can provide an extremely rapid service at a low cost.

The outstanding advantage of machine translation is its speed.

The defects of which users are most aware are due to a variety of causes : those responsible for the system can only hope to improve it in a limited number of areas so as to eliminate the major defects. However, it should be noted that more than 1/3 of the errors pointed out as being the most inconvenient could be eliminated by appropriate pre-editing.

3.2 Analysis of results by type of document

In the following four Tables, the 20 documents in the sample have been sorted by type in accordance with the characteristics listed in § 2.12, so as to calculate the average intelligibility and post-editing rate of their machine translation.

The averages have been calculated both with and without document D 16, the intelligibility of which is exceptionally bad, so as to eliminate the possible impact of an exceptional document.

The figures for intelligibility are taken from the 'average, scale out of 3' section in table § 3.121 (version : automatic translation).

The rates of post-editing are those of version 5 (the only complete one), in Table 3.141.

3.21 Summary table of intelligibility and post-editing rate, by source

Source : Commission			Source : non-Commission		
Doc. No	Intelligibility	Rate of revision	Doc. No	Intelligibility	Rate of post-editing
1	87	36			
			2	92	42
3	88	35			
			4	86	36
5	78	32			
			6	81	26
7	86	34			
			8	73	38
9	74	37			
			10	66	36
11	79	39			
			12	73	34
13	84	41			
14	80	51			
			15	81	28
			16	29	52
			17	80	35
			18	62	56
			19	79	26
20	82	32			
TOTAL	738	337		802	409
excluding 16				773	357
AVERAGE	82.0	37.4		72.9	37.2
excluding 16				77.3	35.7

Comments

- The difference in intelligibility between the two types of documents seems relatively substantial if document 16 is taken into account and relatively insignificant if this document is eliminated.

- There is no significant difference as regards the amount of post-editing.

3.22 Summary table of intelligibility and post-editing rate by document type

ABSTRACTS			UNPUBLISHED TEXTS			PUBLISHED TEXTS		
Doc. No	Intelligibility	Post-editing	Doc. No	Intelligibility	Post-editing	Doc. No	Intelligibility	Post-editing
			1	87	36			
2	92	42						
						3	88	35
4	86	36						
						5	78	32
6	81	26						
			7	86	34			
8	73	38						
						9	74	37
10	66	36						
			11	79	39			
12	73	34						
						13	84	41
			14	80	51			
						15	81	28
16	29	52						
						17	80	35
18	62	56						
						19	79	26
						20	82	32
TOTAL	562	320		332	160		646	266
excluding 16	533	268						
AVERAGE	70.3	40		83.0	40		80.8	33.2
excluding 16	76.1	38.3						

Comments

- Abstracts are less well translated than primary texts, markedly so if document 16 is taken into account and less so if it is excluded
- The intelligibility of the primary documents is practically identical whether they are published or not
- The amount of post-editing is lower for published texts than for abstracts and unpublished primary texts. Can it therefore be said that greater care is given to the drafting of the original texts ?

3.23 Summary table of intelligibility and the post-editing rate, by subject

Agricultural policy			Agricultural regulations			Economic trends in agriculture			Agricultural technology			Food technology		
DOC	I	PE	DOC	I	PE	DOC	I	PE	DOC	I	PE	DOC	I	PE
1	87	36												
												2	92	42
						3	88	35						
												4	86	36
5	78	32												
												6	81	26
7	86	34												
									8	73	38			
			9	74	37									
									10	66	36			
			11	79	39									
												12	73	34
						13	84	41						
			14	80	51									
									15	81	28			
						16	29	52						
												17	80	35
						18	62	56						
												19	79	26
			20	82	32									
TOTAL	251	102		315	159		263	184		220	102		491	199
excluding 16							234	132						
AVER-AGE	83.7	34		78.8	39.8		65.8	46		73.3	34		81.8	33.2
excluding 16							78	44						

Comments

- The subject covered within the general field of agriculture does not appear to exercise any decisive influence on the quality of the system, if badly translated abstracts are eliminated (D 10 - 16 and 18)

- The variation in the amount of post-editing does not appear to be linked to average intelligibility by document type.

3.24 Summary table of intelligibility and post-editing rate, by original language

Likely to be English			Other languages (+) or probably other languages			Language of text unknown		
DOC	I	PE	DOC	I	PE	DOC	I	PE
						1	87	36
2	92	42						
						3	88	35
4	86	36						
						5	78	32
6	81	26						
						7	86	34
			8	73	38			
						9	74	37
			10	66	36			
						11	79	39
			12	73	34			
			13	84	41			
						14	80	51
			15	81	28			
			16	29	52			
			17	80	35			
			18	62	56			
			19	79	26			
20	82	32						
TOTAL	341	136		627	346		572	264
excluding 16				598	294			
AVER-AGE	85.3	34		69.7	38.4		81.7	37.7
excluding 16				74.8	36.8			

(+) including American English

Comments

- The mother tongue of the author of the original document would appear to constitute the only decisive typological factor among those studied : the intelligibility of documents where the language of the text was English or likely to be English attains a level (85.3 %) achieved by no other batch of documents in the three type analyses in § 3.21 to 3.24.

- It should, however, be noted that there is a high proportion of documents where the original source language was not known with any certainty.

- In any future evaluation, it might therefore be desirable to collect data on the source language in a more systematic fashion when selecting the sample of texts for translation.

- If the trend detected here were to be confirmed, we would have arrived at a conclusion with important repercussions for the organization of machine translation : authors of documents to be translated would have to be requested to write in their own language (provided that it existed as a source language in the automatic translation system) rather than in a language they know less well.

- If this trend were confirmed, it would also have a profound influence on the strategy to be adopted in improving the translation system : a simple organizational measure (to request officials of the Commission to write in their own language, rather than in that of their reader) could result in the same rate of intelligibility (85 %) as we anticipate from the next SYSTRAN improvement contract (80 to 85 % : cf. § 1.73)

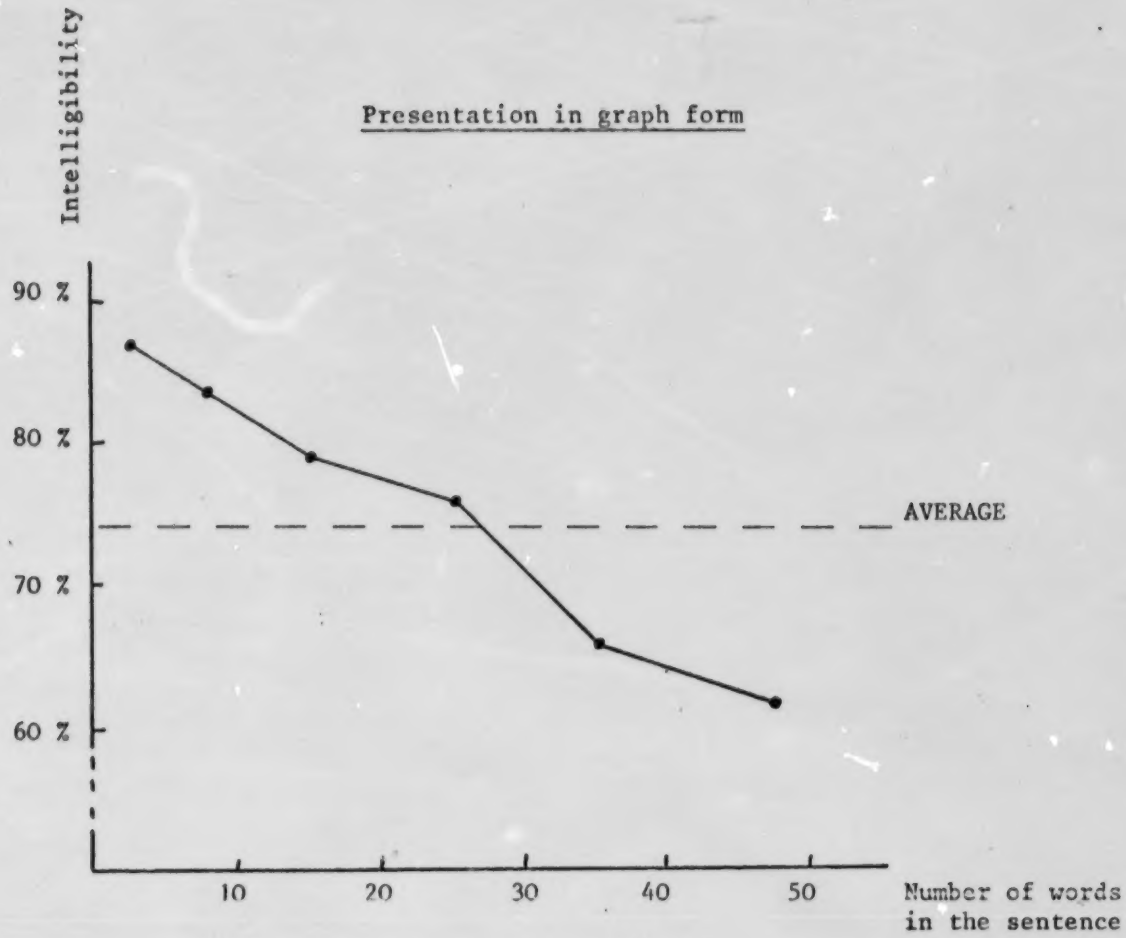
- The differences in the amounts of post-editing required are less significant than the differences in intelligibility; this confirms our previous comment (§ 3.14) as to the relative independence of intelligibility and revision within what is doubtless a fairly wide quality band.

3.3 Analysis by length of sentences

Doc. V 2 No	NUMBER OF SENTENCES, with number of words between							AVERAGE INTELLIGIBILITY (E 8) in % OF SENTENCES, with number of words between						
	1-5	6-10	11-20	21-30	31-40	> 40 (+)	TOTAL	1-5	6-10	11-20	21-30	31-40	> 40 (+)	TOTAL
1	2	5	7	8	10	5	37	100	100	95	88	77	67	86
2	2	3	4	1	3	3	16	100	100	92	100	89	89	94
3	2	6	10	13	9	5	45	100	94	97	82	81	80	87
4	3	2	5	5	1	2	18	78	100	67	80	100	67	78
5	3	3	8	8	6	8	36	44	78	96	88	67	67	77
6	1	2	7	4	4	1	19	100	83	81	67	58	67	74
7	20	7	18	14	2	4	65	100	81	85	76	100	67	87
8	6	2	6	2	2	1	19	100	100	72	83	17	67	79
9	9	2	13	12	6	4	46	85	83	69	81	39	50	70
10	4	3	5	3	2	2	20	83	89	61	67	100	50	73
11	9	10	28	13	4	-	64	89	80	92	64	75	-	83
12	3	-	7	1	5	2	18	100	-	71	100	53	83	74
13	4	5	10	9	8	6	42	75	80	90	85	58	56	75
14	1	3	6	5	1	-	16	100	44	67	73	67	-	67
15	3	3	5	12	9	6	38	100	67	67	94	78	50	78
16	1	-	1	6	4	4	16	0	-	67	22	8	25	21
17	8	8	14	13	9	5	57	79	75	86	69	67	60	75
18	2	1	5	7	4	-	19	100	67	60	52	33	-	56
19	2	-	16	14	5	2	39	100	-	79	81	73	50	79
20	3	1	6	8	5	3	26	100	100	83	63	73	56	74
TOTAL	88	66	182	158	99	63	656	1,733	1,421	1,577	1,515	1,313	1,051	1,487
%	13	10	28	24	15	10	100							
ARITHMETIC AVERAGE								86.65	83.58	78.85	75.75	65.65	61.82	74.35

(+) The average number of words in sentences of more than 40 words is 48.

Presentation in graph form



Comments

- The intelligibility curve is virtually a straight line which can be represented by the following equation :

$$\text{Intelligibility} = 88 - 0.56\ \% \ x$$

where x is the number of words in the sentence.

- The conclusion to be drawn from this phenomenon is that authors should be recommended not to exceed 25 words per French sentence translated from English, that is 22 words per original English sentence, in order to substantially increase the intelligibility of the translation.

4. Annexes

The following documents have been appended :

4.1 Twenty Tables I, containing the analytical results (sentence by sentence) of the evaluation of the 20 documents in the sample.

Each Table gives the results of the evaluation of the 7 versions of one document. Since each version contains a different number of sentences, the following conventions have been adopted to facilitate comparisons between the same sentences in the various versions :

- the sentences in the original version have been numbered in order, starting with 1; the same numbering is used on the text published in 'working documents No 1'

- when 2 English sentences have been translated by a single French sentence in another version, the French sentence is identified by the numbers of the English sentences bracketed together

- where one English sentence has been translated by two sentences in French, the first of these is identified by the number of the English sentence and the second by the same number followed by the 'sign

  Thus, the original numbering of the English sentences has been preserved in the six other versions.

  For even greater safety, each sentence in each version has been identified by the initials of the first three words.

4.2 Twenty Tables II, containing the analytical results (sertence by sentence) of the evaluation of version 2 (machine translation) of the 20 documents.

Note

1) The original texts of the 7 versions of the 20 documents in the sample and of the single version of the 2 control documents, in a typed fair copy, as they were examined by the evaluators, have been published separately under the title 'Working documents Nc 1'. To make comparison easier, the 7 versions of each of the pages of the 20 documents have been presented on a single sheet after photoreduction. A limited number of copies on a 1/1 scale are available at the Commission but the 7 versions appear in separate sections.

2) The computer printout of the machine translations of the 20 documents with the handwritten corrections made by the 3 post-editors (V 5 - 6 - 7) have been published separately under the title 'Working documents No 2'.
No 1/1 scale copies are available.

Tables omitted as unreproducible.